EUROPEAN CENTRE
FOR ELECTORAL SUPPORT

PROSES

Project in Support of Enhanced Sustainability and Electoral Integrity in Afghanistan

Guidelines on Data Analysis and Statistical Forensic Methods to Detect Electoral Irregularities

This project is fully funded by the European Union

European Union

This publication was developed under:
The European Union funded “Project in Support
of Enhanced Sustainability, and Electoral Integrity
in Afghanistan (PROSES)/ECES”

Afghanistan, Kabul 2020
http://democracy-support.eu/afghanistan/

Table of contents

About the Author

The author of this guide, Lukasz Widła-Domoradzki, has 21 years of experience in analysing data and teaching statistical techniques. Lukasz is an experienced user of advanced statistical programmes such as R, SmartPLS and SPSS AMOS.

He has worked for several international research and consulting companies as well as for governmental organizations in Poland. Throughout his career, he has authored over 20 scientific articles about statistics. Lukasz was a leading expert in 2016 evaluation of Polish Aid in Kenya and worked as a Data Analyst during EU EOMs in Liberia (2017), Sierra Leone (2018), Pakistan (2018) and Nigeria (2019). In Afghanistan, he provided training in data analysis for election stakeholders, under the EU-funded project PROSES, implemented by ECES.

About the Project

The European Centre for Electoral Support (ECES) is a not-for-profit foundation headquartered in Brussels which promotes sustainable democratic development through the provision of advisory services, operational support and management of large projects in the electoral and democracy assistance. ECES is implementing the Project in Support of Enhanced Sustainability and Electoral Integrity in Afghanistan (PROSES) funded by the European Union through the Instrument contributing to Peace and Stability (IcSP). PROSES outcomes are:

Outcome 1: Electoral integrity vulnerabilities are proactively identified and reduced through administrative planning and actions.

Outcome 2: Increased capacity of electoral stakeholders to apply evidence-based and effective leadership approaches to the conduct of elections and the adjudication of election disputes.

Outcome 3: Measures contributing to the accountability and inclusivity of political and electoral processes through the broad participation of relevant electoral stakeholders with specific emphasis on women.

PROSES' activities feed into a comprehensive electoral integrity strategy based on the analysis of integrity vulnerabilities from previous Afghan electoral cycles, recommendations from EU EOM/EATs and an upcoming electoral political analysis. PROSES is supporting vital electoral stakeholders' efforts in promoting enhanced integrity and credibility of the electoral process. In brief, our overall strategy is multi-stakeholder: partnering and supporting all individuals, organisations and processes (civil society, media, political parties, EMBs, Government, justice system) that can play a role as agents for reform

and change in the Afghan electoral process and democracy. To identify and support Afghan-led and Afghan-owned collective actions to mitigate fraud and malpractice through the work of alliances and coalitions with key Afghan players. While the tangible impact of many integrity actions will take time, the strategy intends to lay the foundations, through multiple actions and milestones, to produce changes in the mid-term and long-term by identifying agents for change and meaningful actions.[1]

Goals of these guidelines

The overall goal of these guidelines is to outline some fundamental concepts of electoral data and electoral analysis, followed by practical tips to be used as you build a database using open-source software. These guidelines are composed of four sections. The first setion introduces the basic concepts of electoral data and statistics. The second section provides information about how to collect and build a database. The third section introduces two useful tools, Microsoft Excel and QGIS, to create databases and maps, respectively. The last section focuses on statistical analysis to detect election irregularities and fraud through the analysis of statistic patterns occurred on E-day.

SECTION I

What's electoral data and statistics?

Mathematics and statistical analysis are vital to understanding the processes of everyday life. Modern societies are confronting views with numbers daily: sooner or later everyone comes across a statistical issue explained by a newspaper or on TV, on a public announcement or just by a friend. In the electoral field, data and statistical analysis allow explaining patterns and understanding results, as well as identifying possible irregularities and fraud, by using electoral data gathered on election day or throughout other phases of the electoral cycle. The chances of a given candidate or political party are measured by the percentage of people who are willing to vote for him or her according to surveys and polls. Campaign spending is counted in actual money, but also – in many countries – as a percentage to observe a spending limit. There are speculations about possible voter turnout and predictions on what level of turnout will bring what outcomes. All these numbers are produced using statistical techniques. Yet, data analytics plays a more important role in elections than just predicting the outcome. Analytics is also increasingly becoming an integral part of political campaigns. In recent elections across the globe, political parties have employed data-driven analytics and social-media data to stay ahead of the competition.

[1] Further read on electoral integrity can be found at http://www.eods.eu/publications

Uses of Electoral Statistics

In the electoral context, we can use statistics for three main purposes:

1) Predictions
2) Factual analysis
3) Detection of electoral. irregularities and fraud

Usually, **predictions** are based on incomplete data. Before all the results are announced, a statistician can extrapolate available data. Extrapolation of incomplete data is a complicated and complex process, and the analyst should be aware of the nature of the non-available data, usually using other databases or expert knowledge from other sources. Due to its complexity, this document does not cover the prediction role of statistics.

Factual analysis is useful when preparing an analysis of the observed reality. For instance, factual analysis is used for the calculation of turnout figures after the elections. Another example of factual analysis is voting patterns maps. Factual analysis is the representation of the reality through maps, charts, etc. While this paper does not cover factual analysis per se, it will explain the potential use of an adequate databases for data analytics.

Fraud detection is similar to factual analysis, but instead of description of the reality, fraud detection is focused on identifying discrepancies. To perform fraud detection, a statistician may use only publicly available data (for example to check unusually high or low turnout or inexplicable change of voters' sympathy in a given region) or compare different databases. The second condition is possible only if other (unofficial) databases exist – some entities (like observer groups or party agents) are collecting evidence from the field during election day. This intelligence can be transformed into the database and used in some analysis to detect fraud. Fraud detection is precisely the focus of this paper.

Why we need electoral data?

There are different reasons for collecting electoral data, primarily because we need to know who won the elections. But different stakeholders will have different needs regarding electoral data. This kind of information may be collected to:

- … have an accurate view/control of the electoral cycle across 34 provinces and over 400 districts;
- … take informed decisions about the management of the process and the management of IEC staff;
- … keep a historic track of the Afghan electoral process across cycles
- … produce reports, maps, graphics for internal use – professionalism and good management;

- … produce reports, maps, graphics for external use – transparency and accountability towards political parties, observers, donors, and voters.

Most of the needed data will be prepared by the IEC in Afghanistan, but some data may be collected by other electoral stakeholders – party agents, NGOs, or the media.

Human resources

Having a proper database is the basis for data analytics. However, **it's good to have someone who can do the analysis** and interpretation. The analysis is the processing of the data, while the interpretation is applying it to the context. To accomplish this-, you will need an expert on numbers and analysis. It doesn't necessarily have to be a statistician, but definitely someone who will be able to perform a simple analysis and tell if the tools (forms or questionnaires) are properly built. **The person who is responsible for analysis is very rarely the same person who is doing the interpretation**: I may be an expert in the statistics, but know barely anything about Afghan electoral system, recent law changes or types of media present in the country.

Sources of electoral data

As mentioned before, **the main source of electoral data is – and should be – IEC.**

However, there are other sources than can be useful:

- **Party agents**. Each party agent has the right to get a copy of the results. Copies can be transformed into data
- **International Observers**. Usually limited in number, but observations, if conducted properly, generate a data of patterns, irregularities and overall assessment
- **Media**. The role of media is supplementary to collecting data but crucial to disseminate it. Media can also produce visualisation and analysis
- Last, but not least – **NGOs and citizens**. Every single voter has the right to collect and analyse data.

With many different sources of electoral data, the information gathered can be cross-validated and triangulated. Not every source will be as reliable as the rest, but with many sources, those of them that are not trustworthy should be easily distinguished.

The databases gathered by political parties are oftentimes the least reliable. As political parties are obviously biased in favour of their party or candidate. It does not necessarily mean that all the data gathered by the political parties are useless – if more than one party is showing the same result, the data is probably reflecting reality.

On the other hand, if you are a citizen or an NGO worker, you know already how hard is to get a reliable database with sufficient data points (observations). A single citizen can't gather all the raw data throughout the country and it's next to impossible to get this kind of information for smaller areas (like cities or districts). It's not an easy task, even if you are an NGO worker – even if your organisation is big or it is a part of the NGO association. The main advantage of gathering data by NGOs is their perceived impartiality. The crucial condition here is the proper training of those who will be collecting data in the field: people may have different backgrounds or ideas about what data is important to collect. Good quality and impartial data are essential for preserving the integrity of any democratic process.

Collecting and disseminating sex-disaggregated data is a key task for inclusive election management, since it allows EMBs to assess gender balance in the electoral process and to better plan their strategies and policies. However, data on the gender composition of registered voters, voters who actually casted their votes on Election Day, registered candidates and electoral staff at all levels is not always available. Data gathering design and processes should consider the gender dimension from the beginning, including a "sex" or "gender" category as one of the data fields and recording it at the time of data collection. Through training initiatives, EMBs can strengthen the capacities of staff and elected officials to collect sex-disaggregated data

SECTION II

Basics to build a statistical database

Without data, you're just another person with an opinion – W. Edwards Deming.

Gathering and using data is crucial today. Using elections data to perform analysis and interpretation of trends and exceptions can provide useful information to identify irregularities and possible fraud. When collecting data, the database has to be both orderly and complete.

- **Orderly**. Every single question should be written down as a numerical entry that can be used to create analysis and graphs. Orderly also means "neat": a database should be easy to use, well labelled and transparent to the people who will work with it.
- **Complete**. To perform effective statistical analysis, as big a database as possible is needed. Having a proper database is the first and most important step in gathering information.

If you are the organisation who will be gathering information from the field and you would like to build a statistical database, you should be aware of several things. First, every entry in your database will be probably collected separately. Your role is to put all of the entries in one database. Scattered information is – technically speaking – a database, but it's useless from the statistical point of view. You will be able to compare your findings with official data only if you prepare a complete and ordered database.

In other words, a statistical database should be usable. Electoral data should be easily verifiable to all the stakeholders (including citizens). That's why there is a need to have the data in a usable format (e.g. Excel spreadsheet). Publishing database in a not usable format (e.g. if the form of scanned Polling Stations results sheets) can only make the whole electoral data analysis process harder – different electoral stakeholders will be able to use this information only after prescribing the data into the proper numeric format – which consumes time, money, and workforce.

Secondly, before getting any answers from the field, you will need a checklist of all the topics you'd like to address with your database. This checklist should contain all the questions you would like to answer after the elections. In sociology, this part of the work is called "hypothesis testing". Your checklist may contain specific questions (e.g. "was the Polling Station opened at xx hour?") or broader categories (e.g. "assessment of security forces"). Each organisation has their own way to deal with checklists.

If you don't specify something in the checklist, you won't be able to answer the underlying question afterwards. In other words, a database will be only as good as is the checklist

Last, but not least, you will have to answer some questions to yourselves, before any data collection starts. Oftentimes, election forms (or questionnaires) contain too many questions. It's your part of the work to ask: "will I really use this question afterwards?". If the answer is negative, you better spare your time and the time of your people in the field and remove unnecessary questions from the form.

Your statistical database may contain:

- Observation forms
- Complaints
- Information about political parties' activity
- Previous elections results, etc.

There are many advantages to keep the statistical database in a numeric readable format (and not as – for example – pdf file):

- You can change figures when needed
- You can use it as a basis for statistical analysis
- You can prepare graphs and maps to communicate the results

Database building is complicated and it's divided into several steps. From the first thought about building database, to conceptualisation (checklist) and observer training to the selection of the program for data collection – the road to the statistical database is long.

Analysis with electoral data

Your analysis, based on electoral data should be clear, credible/consistent and accurate:

- **Clearness**. Figures used for presenting electoral data should be self-explanatory. Only then they are verifiable. First, the basic numbers should be announced: for example, everyone will be able to check the turnout figure if two numbers will be known:
    - total number of votes
    - total number of eligible voters

  It's advisable to say: "During the election xx voters voted and at the register roll we have yy registered voters. Xx divided by yy is zz%, therefore this is our turnout figure" – the argumentation here is **clear**.

- **Credible, consistent and accurate**. Electoral data must reflect the reality! Electoral data should be free of any flaws, discrepancies and errors. However, errors, if present, should be easily explainable.

It is worth mentioning that your information will be probably a collection of both qualitative and quantitative data. Working with statistics means using only numeric data which implies coding all qualitative data. More on this subject below.

SECTION III

Analytical and Graphic Tools

To detect irregularities and fraud, but also trends and patterns, appropriate statistical tools need to be used. In most cases, simple tools are sufficient, for example, Microsoft Excel or LibreOffice Calc. In special cases (e.g. time series analysis) specific statistical packages can be used. The most common are R, SPSS, Stata, Statistica and SAS.

Microsoft Excel

Excel is the ideal tool to build a statistical database. A statistical database may contain:

- information about candidates' activities (rallies, incidents, etc.)
- turnout per polling station or polling centre at district, province or national level
- preliminary or final results (ongoing elections, previous elections)
- data from national or international election observers
- complaint by types (suspicions of fraud, etc.)

Excel chart (below) - Breakdown of political parties' agents presence per party in a selection of polling stations at different stages of Election Day (opening, voting, closing, counting).

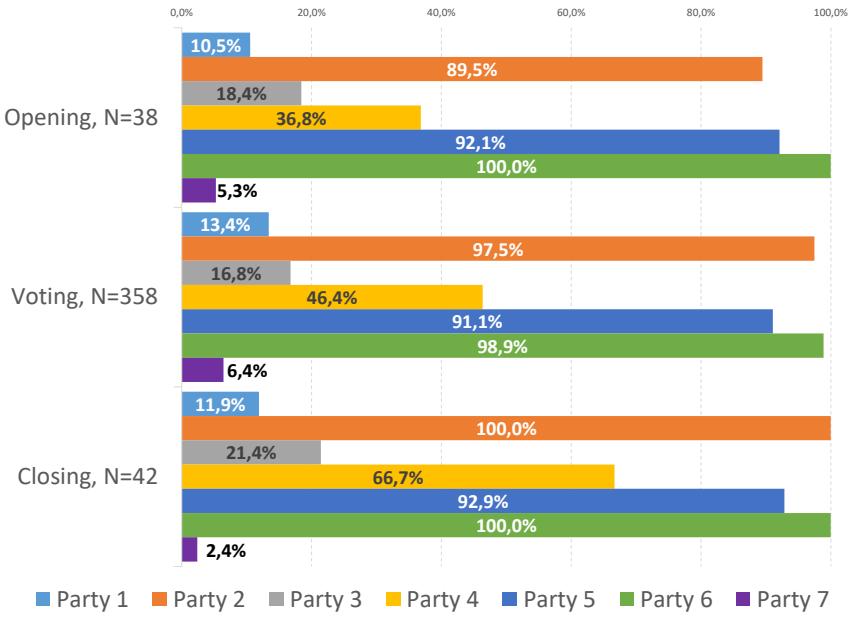


Table. 1. Example of a raw database

Form B Voting											
Source	Username	A.2 Time of arrival	B.2A Polling station constituency number (3 digits)	B.2B Polling station ward number (3 digits)	B.2C Polling centre code (4 or 5 digits)	B.2D Poling Station number (2 digits)	B.3 Polling station type	B.4 Was this polling station chosen randomly?	C.1 Is the PS accessible for voters with reduced mobility?	C.2 Is there a long queue (more than 30) of voters waiting to vote outside the polling station?	C.3 Did you observe any other problems in the vicinity (within 400 m) of the PS?
app	Itosl	07:05	4	10	1070	2	Urban	Yes	Yes	Yes	No
app	Itosl	07:00	50	169	7074	1	Rural	Yes	Yes	Yes	No
app	Stosl	07:16	50	170	7088	2	Urban	No	Yes	Yes	No
app	Itosl	07:21	128	436	16199	6	Urban	Yes	Yes	Yes	No

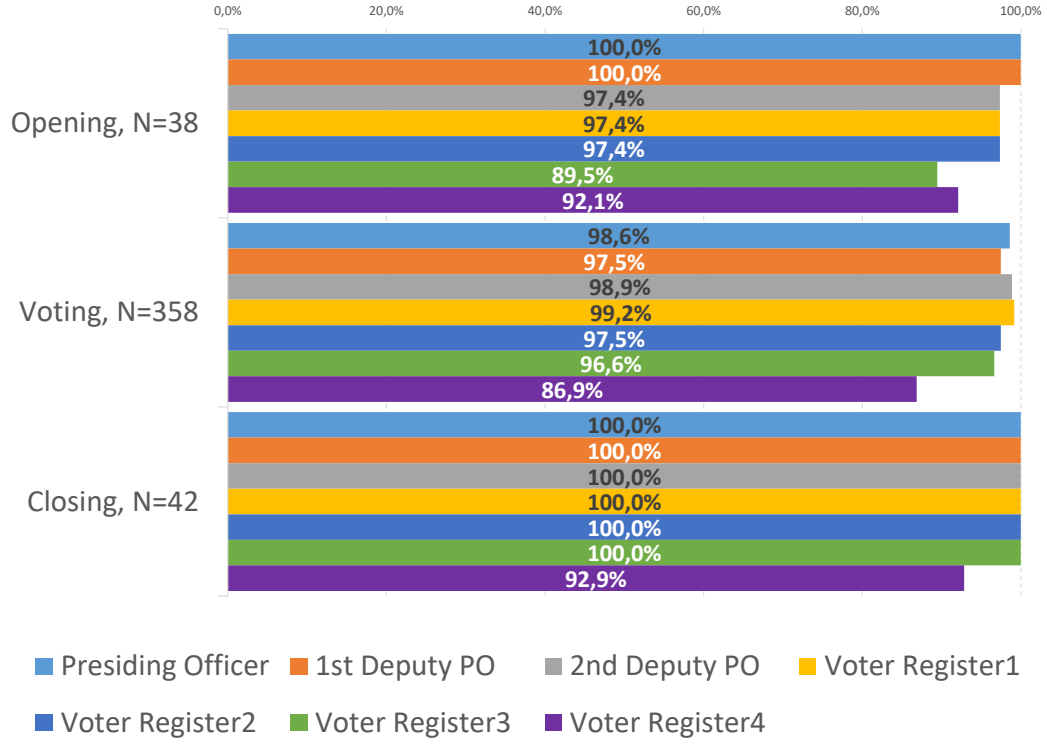
This is an example of a raw database. Raw means “just as it came from the system”. Your database may look different because you might use a different system. In the raw database, you will find numbers or text answers – those need to be recoded to the number format:

Table 2. Sample transformed database

Source	Username	A.2 Time of arrival	B.2A Polling station constituency number (3 digits)	B.2B Polling station ward number (3 digits)	B.2C Polling centre code (4 or 5 digits)	B.2D Poling Station number	B.3 Polling station type	B.4 Was this polling station chosen randomly?	C.1 Is the PS accessible for voters with reduced mobility?	C.2 Is there a long queue (more than 30) of voters waiting to vote outside the polling station?	C.3 Did you observe any other problems in the vicinity (within 400 m) of the PS?	C.3.1 If Yes, please specify:	C.3.1 If Yes, please specify: (Campaign material
app	Itosl	07:05	4	10	1070	2	Urban	1	1	1	0		
app	Itosl	07:00	50	169	7074	1	Rural	1	1	1	0		
app	Stosl	07:16	50	170	7088	2	Urban	0	1	1	0		
app	Itosl	07:21	128	436	16199	6	Urban	1	1	1	0		
app	Stosl	07:36	75	235	10017	2	Urban	1	0	1	0		
app	lstosl	07:33	124	428	16150	3	Urban	1	1	1	1		
app	Itosl	07:19	96	338	13110	1	Urban	1	0	1	1		Campaign material
app	Stosl	07:15	75	235	10014	2	Urban	1	0	1	0		
app	Itosl	07:21	32	106	4051	3	Urban	1	0	1	0		
app	Stosl	07:44	44	150	6055	3	Rural	1	0	1	0		
app	Stosl	07:45	113	399	16006	4	Urban	1	0	1	0		

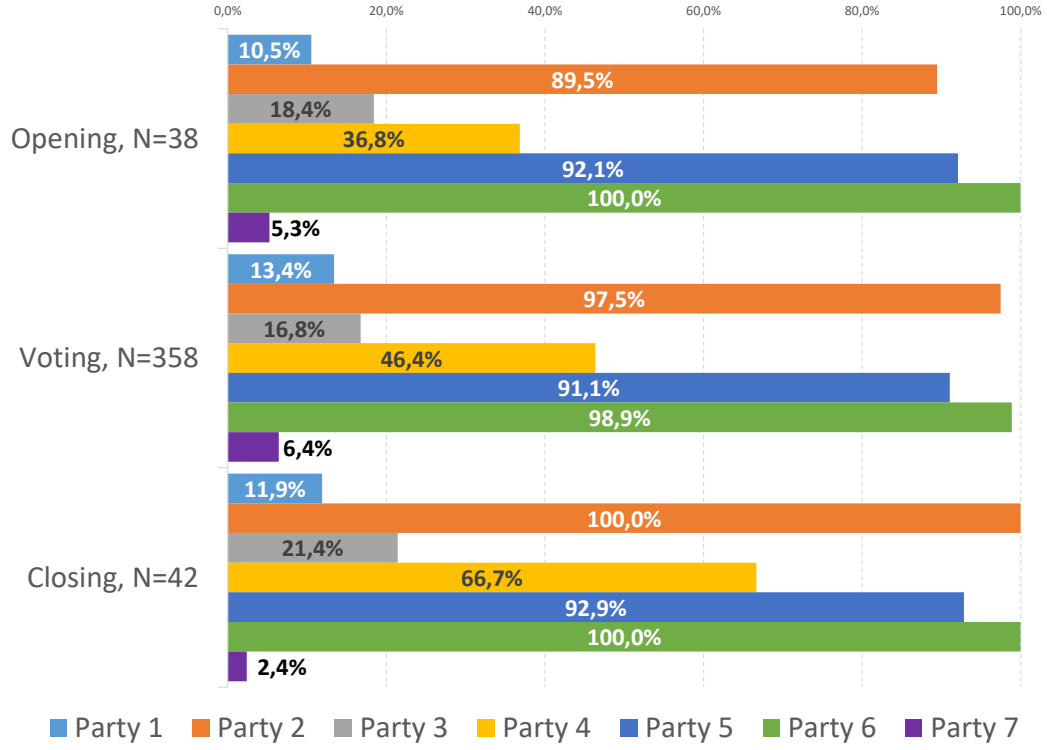
As you can see, some elements from the raw database disappeared: there is no information about the type of the form used (see row 1 from Table 1). Also, some of the texts were replaced with numbers. At this example data from the original database from columns I:L was replaced with 0/1 coding. In this case “0” represents “no”, while “1” represents “yes”. This kind of transformation will allow you to prepare charts.

Chart 1a. Sample chart “People present inside the Polling Station”



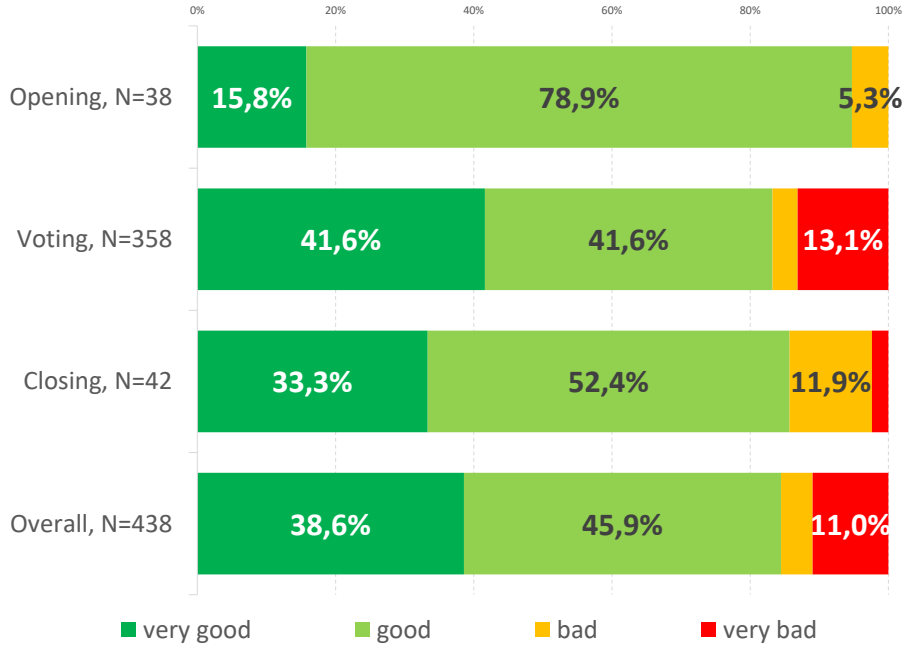
Source: own calculation. “N” at the chart represents the number of observations

Chart 1b. Sample chart “Party representatives present inside the Polling Station”



Source: own calculation. “N” at the chart represents the number of observations

Chart 1c. Sample chart “Overall evaluation of the Election Day”



Source: own calculation. “N” at the chart represents the number of observations

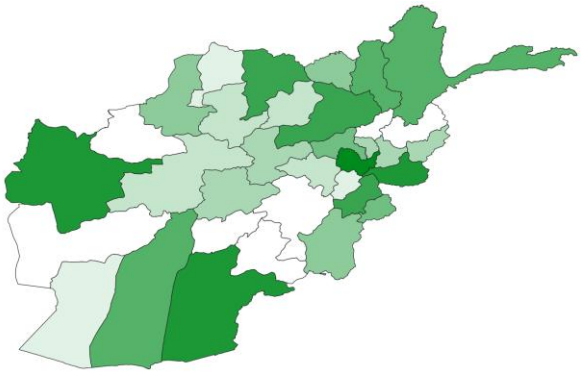
GIS (QGIS) analysis

For GIS analysis, use of QGIS tool is recommended[2]. This tool is helpful to check spatial (geographical) patterns. For this analysis, a shapefile of the country is needed (for example from DIVA-GIS: https://www.diva-gis.org/gdata)

First step to create a map is to get “shapefile”

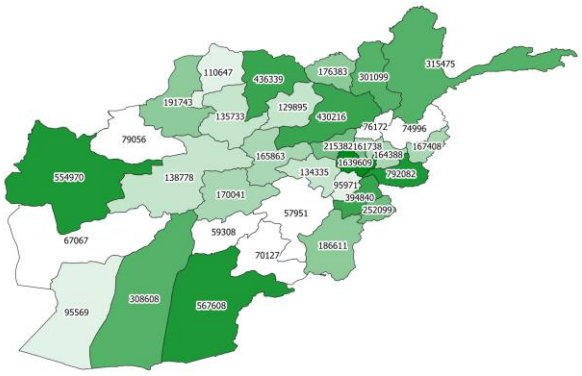
The shapefile (usually a file with .shp extension) is the blank map with no data at all – just a shape of a country and a desired division. Usually the data available is the whole country shapefile, district shapefile, province shapefile, but it depends on the country.

[2] https://qgis.org/en/site/



Registered voters in Afghanistan (2018)

Here is the map for the registered voters in Afghanistan. This map is prepared using an organized numeric database (here we used information from the excel spreadsheet file). Remember, to prepare a GIS map you need a shapefile and the numeric data you would like to put on the map.



Registered voters in Afghanistan (2018), with numbers

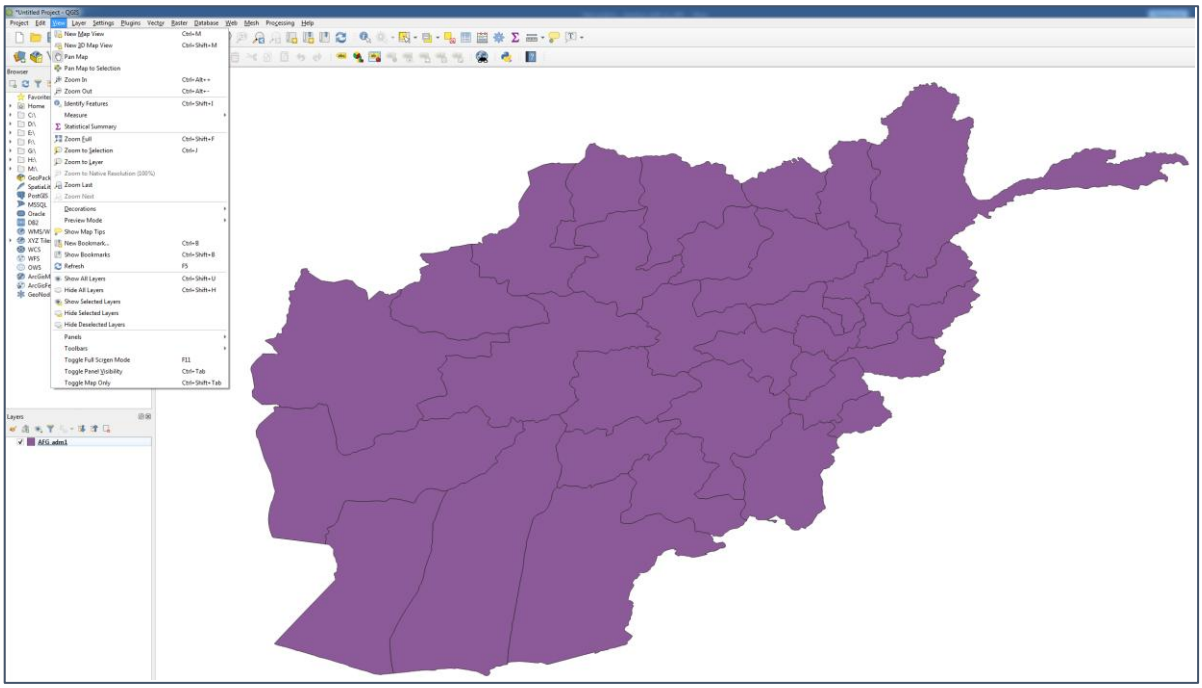
Here is the same map, but with number tags. Sometimes it's preferable not to have all the numbers visible on the map as in this one the labels for Parwan and Kapisa provinces overlap.

QGIS – Geographical Information System Tool

Sometimes, simple percentages in the graph are not sufficient. It's good to know basic statistics, and in some cases it's worth taking a look into geographical patterns. To do this the use of GIS programs is arguably the best solution. I will show you at the Analysis section how to use QGIS for searching "strange" voting patterns and how QGIS can facilitate showing a pattern which will be not visible in a chart or graph.

Among various GIS programs (the most popular, but also costly, is ArcGIS) QGIS is probably the most advanced free of charge tool. It's quite complex as an open source tool, yet quite easy to operate if you need is to prepare a simple map.

Picture 1. QGIS user interface

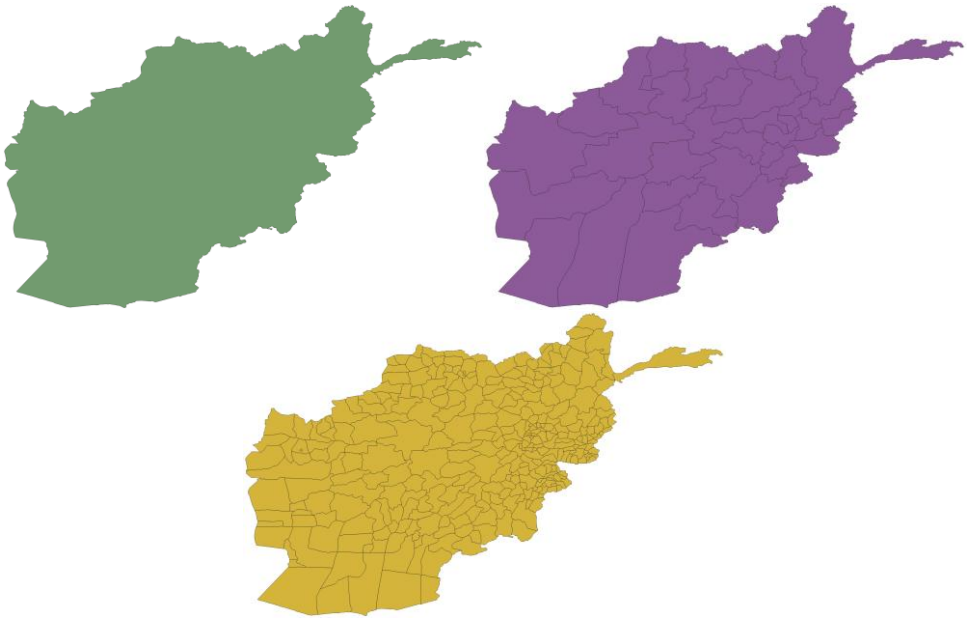


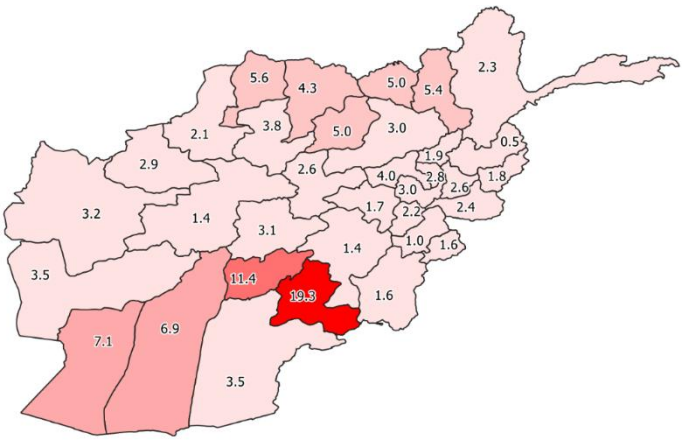
To prepare a map using a QGIS you'll need

- The software itself
- A shapefile of the country you'd like to map

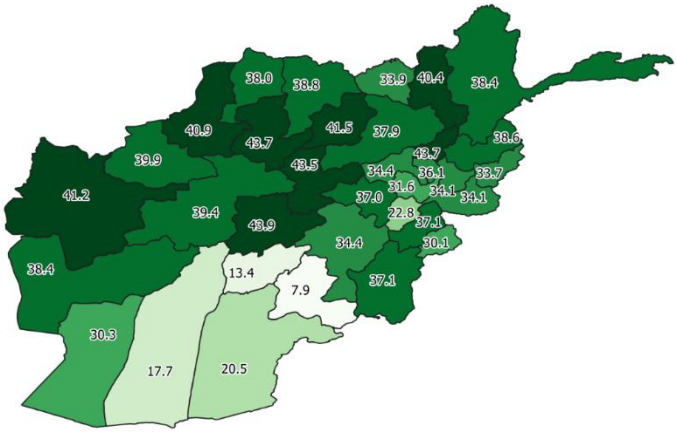
A shapefile is an outline of the country with different divisions. Each of three maps shown below are separate shapefiles:

Sample map 1. Afghanistan shape files





Sample map 2. Percentage of invalid votes (2014 Afghan election data)



Sample map 3. Turnout of the female voters (2014 Afghan election data)

Shape files, as well as QGIS can be obtained for free from different Internet websites. Possible sources are:

For QGIS program:

https://www.qgis.org/en/site/forusers/download.html

For the shapefiles:

https://www.diva-gis.org/gdata

The guidelines on how to install the QGIS can be found in the ANNEX.

SECTION IV

Statistical Analysis

After getting and installing the right tools, we can move to the analysis section. Rightfully chosen analysis can discover irregularities and fraud as well as show patterns and consistency. Below you can find some examples of the application of various methods of statistical analysis to detect electoral irregularities.

Comparative analysis

By using more than one database (e.g. from current and previous elections), it is possible to compare voting patterns. Voters' sympathies may change, but the assumption is that in similar regions the sympathy for a party should change in a similar way. If the results are showing a different pattern, the researcher will have to look for the reason why. This does not necessarily mean fraud has occurred, but further investigation may be warranted.

Comparative analysis is a branch of statistics. Basically, it's the process of comparing information from different sources or from the same source, but at the different times.

One of many possible applications of comparative analysis can be detecting some irregularities in the election results. For example, let's assume that Candidate A got 40% in the first round and the Candidate B got 25% in the first round.

Probable outcome	**Remarks**
Usually it is more likely for Candidate A to maintain the winning position. His / her voters will not disappear for the run-off and it may be easier for him / her to convince more people to vote on his / her behalf. The rationale here is: - it is easier to maintain a winning position if you have to gain just several more % to win - it is easier to convince people you are the one if you are getting more votes during first round	It may be not true if there is one strong ruling party and scattered opposition: opposition leaders that are no longer compete might ask their voters to vote for the remaining opposition leader. In some countries voting is driven by race, clan, language, ethnics etc. In this case there may be a situation when winning candidate will be unable to get more votes for the run-off – because his / her pool of voters is depleted.

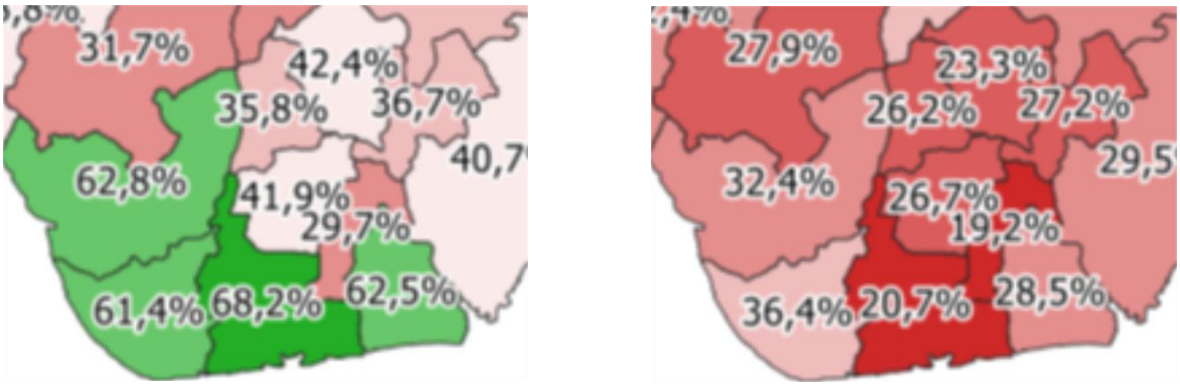
If in the constituency Z Candidate A got 40% and Candidate B got 40%, then you have to look closely at what is going on during the second round – this constituency is sensitive! Sensitive areas/constituencies are easy to spot and they will need special attention during the run-off. The temptation for party agents to change results on behalf of their candidate will be bigger than in non-competitive areas. Some candidates deliberately abandon campaigning at the areas they know they have no chances of winning. Their focus will be at the areas they have a chance of winning– competitive ones.

What we are searching here are not simple anomalies – those may happen in particular polling stations. The scope here is to search – for example – totally inexplicable inversion of the voting patterns. If the given territory was leaning towards a candidate representing a particular point of view or a candidate from specified ethnic group or a candidate whose mother language is x, then we expect similar voting patterns for the run-off.

Turnout and invalid votes analysis

Suspicious turnout rises or fall is one of the tools of messing with election results, but it may be hard to prove if the turnout figures are not reliable.
Below is an example of how turnout may be used to create election ambiguity.

Picture 2. Turnout analysis sample



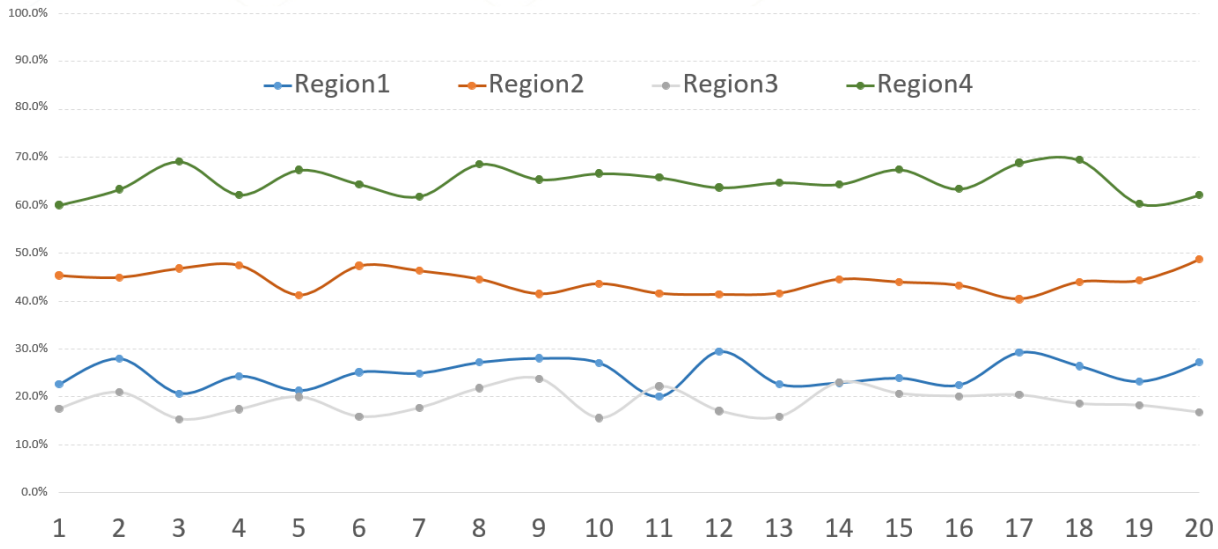
In the left side we can see the turnout for the 2015 elections in the Country X. At this time, Party A was ruling the country and their stronghold was in the south. The overall turnout figure for the country in those elections was around 40%, yet in the stronghold of the ruling party, it was nearly 30% higher.

However, despite of the great result in their stronghold, the ruling Party A had to go to the opposition benches, and former opposition party, Party B, rose to power. Shortly after, Party B nominated the new Election Commissioners. At the right side of the picture we have the situation from 2019 elections – the stronghold for Party A, now opposition party, reports a very low turnout – about half of the overall country result!

Does this mean election fraud? Not necessarily. There are some situations when the opposition, for example, boycott elections. Or, there may be other circumstances that forced the Election Commission to quarantine or even cancel results – excessive violence, breaking the election law, indications of buying votes, ballot box stuffing etc. If the votes (Polling Stations, Polling Centres or even whole Constituencies_ are put under quarantine or are cancelled – it has to be visible for the turnout / invalid votes analysis. In fact, that was the explanation for the exceptionally low turnout in this particular country described above – election-related violent incidents forced the election commission to cancel over one million votes! [3]

Turnout analysis may be conducted also at the polling station level. If in the given region for most of the polling stations the turnout is more or less constant (e.g. 20%-30%) and just for few polling stations is exceptionally high or low – then the researcher should have the opportunity to ask why is that: turnout for similar polling stations in the given region should be similar as well:

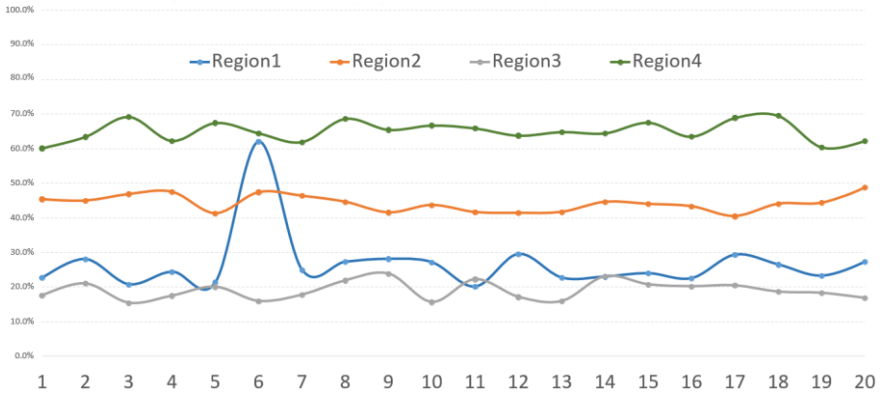
Chart 2a. Turnout at the Polling Stations level sample



Of course, in a country like Afghanistan, several provisions should be met, for example comparing the same type (male or female) of polling station. For cultural reasons, we can expect lower turnout for female polling stations, if someone put at the chart only male polling stations and one female PS among them – the question about why the turnout is not in line with the rest will not be well-grounded:

[3] However, it is questionable if legal issues were the sole reason for cancelling the votes. The turnout analysis may discover in this case the phenomenon of exceptionally low or high turnout. The explanation of this observation may vary: from fraudulent activity to perfectly explainable phenomenon.

Chart 2b. Turnout at the Polling Stations level sample



Results meta-analysis

Once announced, everyone should be able to check and corroborate results. Meta-analysis means we can divide larger parts of the country (provinces to districts, districts to constituencies and so on) and check if the results for larger parts are in line with results for smaller ones. If, for example, in the presidential elections the Electoral Commission announces that in the Province X, Candidate A got 52% of the votes and your observers reported from the field that in all observed polling stations Candidate A got no more than 20% of the votes – it may indicate:

- Human error. Maybe someone just put 52% instead of 25%. It happens! It is not uncommon to discover 120% turnout with a little help of a simple clerical error. It may happen just because someone switched numbers and put 540'000 valid votes instead of 450'000.

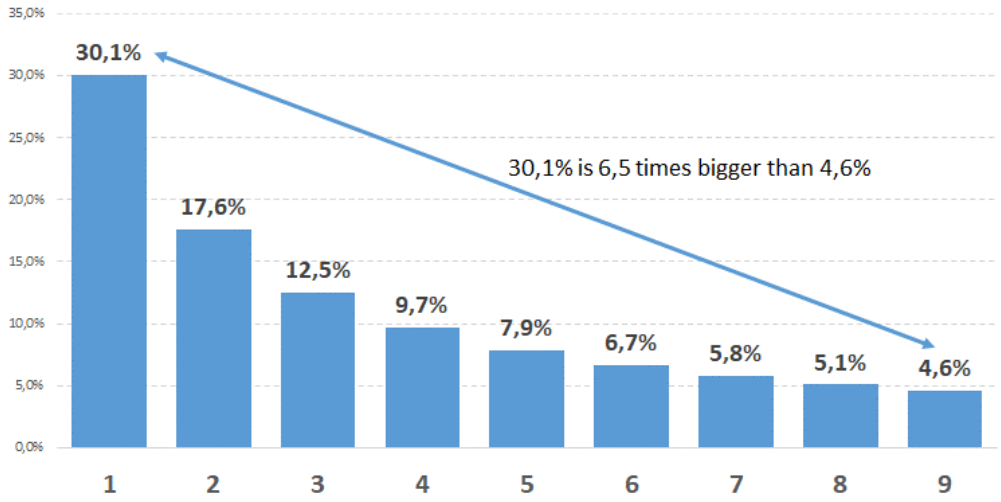
- There may be also a situation when an observer sample is not representative of the larger part or just skewed in another way. It happens when the collected sample is not random[4]. It's important to cover all the different types of polling stations. Maybe women voted differently (and observers observed only male polling stations, or vice-versa)? Maybe there are minorities in the region (and observers observed only constituencies without minorities)? Maybe remote villages voted differently (and observers were present only in the cities)? There are many ways to skew the sample.

- Finally, it may mean fraud. Because if – for example – all the Polling Stations in a given area reported no more than 30% for the candidate A and the announced result for that region is 40% for the candidate A, then something definitely went wrong.

[4] The randomness of the sample is one of the statistical main principles. This principle is rooted in the law of large numbers and the combinatorics science. For more information see (English): https://www.encyclopedia.com/science-and-technology/mathematics/mathematics/randomness

First digit law (Benford's law)

The term 'first digit law' refers to the appearance of the first digits in a given variable taken from a given dataset. In real datasets, there is an expectation that the lower first character – the higher the frequency. In fact, number "1" should appear as a first number 6,5 more frequent than "9". Benford's law is used sometimes to discover election fraud or – more precisely – if there is an indication of human altered information. This tool can be used to analyse large sets of data such as provincial or country elections results sheets.



When people are changing the information, they tend to choose the same numbers with more or less the same frequency. If I put "1" as a first number I'll probably put another "1" after 10 – 12 alterations. As a result, I'll get uniform distribution of first numbers: the expectancy here is to get 11,1% of "1", 11,1% of "2" and so on. Variables changed by humans will – in most cases – broke Benford's law, because our brains are working usually with no prejudices against numbers – for us number "7" as a first digit is as good as "2" as a first digit, while for the Benford's law "7" as a first digit is triple less probable than "2"!

Different versions of Benford's Law were indeed used to detect election fraud. A technical, but interesting paper about competition between George W. Bush and John Kerry in Florida available at 2004 and an example from Mexico elections 2006 can be found here:

https://www.researchgate.net/publication/237341643_Election_Forensics_Vote_Counts_and_Benford's_Law

Steps to detect election fraud using election data

- Use verifiable data rather than rumour or hearsay
- Use proper and complete databases
- Check for unusually high/low turnout in comparison to the previous election
- Check for unexpected rise or fall in turnout between two rounds of the election (it may be at the province/district or even at polling station level)
- Remember that turnout for the bigger area should be generally in line with the turnout for the smaller parts
- Some constituencies, districts or even provinces will be highly competitive. Pay extra attention to them
- Search for "odd" patterns, especially geographical ones
- Search for strange numerical appearances: remember Benford's law of first numbers[5]

Detecting election fraud using numerical forensic methods

There are a number of triggers which can alert analysts of potential fraud and these can also be used by journalists as data to draft graphs and charts for analysis and illustration. These are, among others:

- the number of ballots cast cannot exceed the number of voters having voted
- the maximum number of voters assigned to a polling station is 400
- the total of valid votes, invalid votes and blank ballots has to equal to the number of ballots received before the election
- if the number of votes for a given candidate exceeds a certain high percentage
- if the number of invalid ballots exceeds a certain high percentage
- exceptionally low or high turnout (especially in female polling stations or insecure areas)
- any sensitive material missing/irregularities connected with Tamper Evident Bags (TEB), for example results published not in line with results from TEB

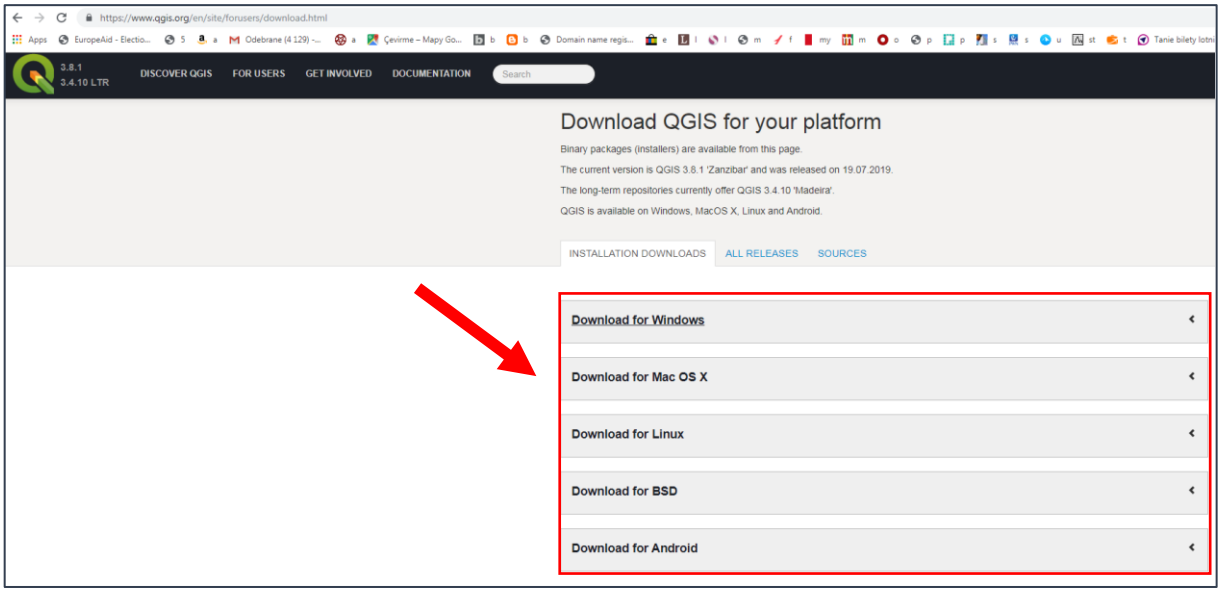
[5] Benford's law (also called the *first digit law*) states that the leading digits in a collection of data set are probably going to be small. For example, most numbers in a set (about 30%) will have a leading digit of 1, when the expected probability is 11.1% (i.e. one out of nine digits). This is followed by about 17.5% starting with a number 2. This is an unexpected phenomenon. In statistical terms, Benford's law is a probability distribution for the likelihood of the first digit in a set of numbers (Frunza, 2015).

Fraud activity	How to check
Ballot box stuffing	Number of ballots inside the ballot box exceeding 400 Percentage for the given candidate is exceptionally high
Unjustified annulation or quarantine	Annulation of votes or quarantine only in specific areas
Abuse of female voting	Suspicious results from female polling stations from specific areas
Use of ghost polling stations	Suspicious results from polling stations from insecure areas (e.g. presumably not opened due to the insecurity)
Forging results at the tabulation centre	Results not in line with the results posted outside the polling stations Results from the given tabulation centre differ from other tabulation centres

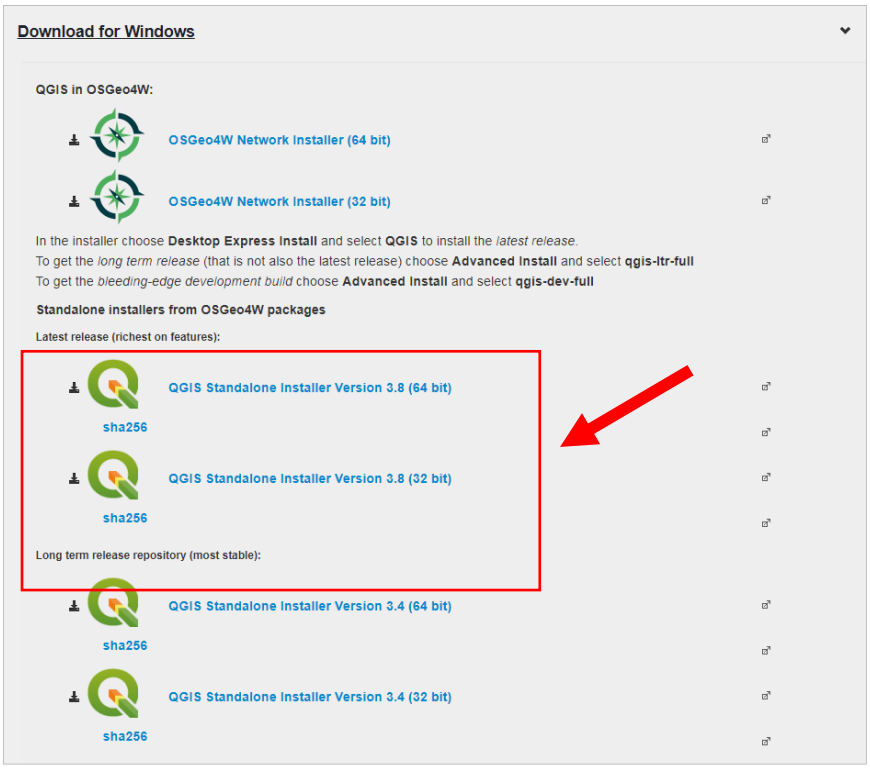
ANNEX

Five steps to install QGIS

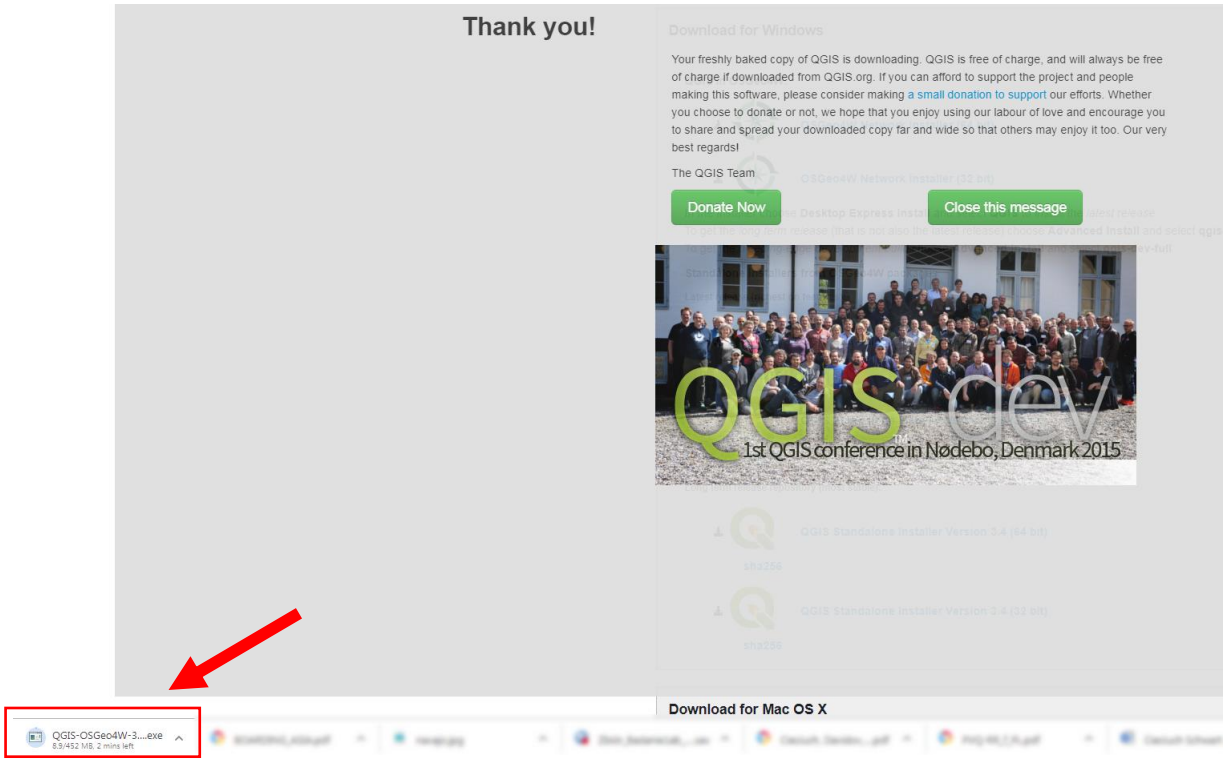
1. From the download site you have to choose your system:



2. If you choose, for example, 'Download for Windows' you'll find multiple options. Choose either of the "QGIS Standalone Installer Version". You should download "64 bit" or "32 bit", it depends on the version of your operating system.



3. The download will start automatically, and you'll be redirected to the "thank you" site.



4. After download start the installer and install the program:

5. You're ready to start using QGIS!

This document has been produced with the financial assistance of the European Union. The contents of the document are the sole responsibility of the European Centre for Electoral Support (ECES) and can under no circumstances be regarded as reflecting the position of the European Union.