Accepted for Publication

This version of the article has been peer-reviewed and accepted for publication in *Scientific Data*. However, this is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available at: doi.org/10.1038/s41597-025-06502-7

The Malaysian Election Corpus (MECo): Federal and State-Level Election Results from 1955 to 2025

Thevesh Thevananthan[1]✉

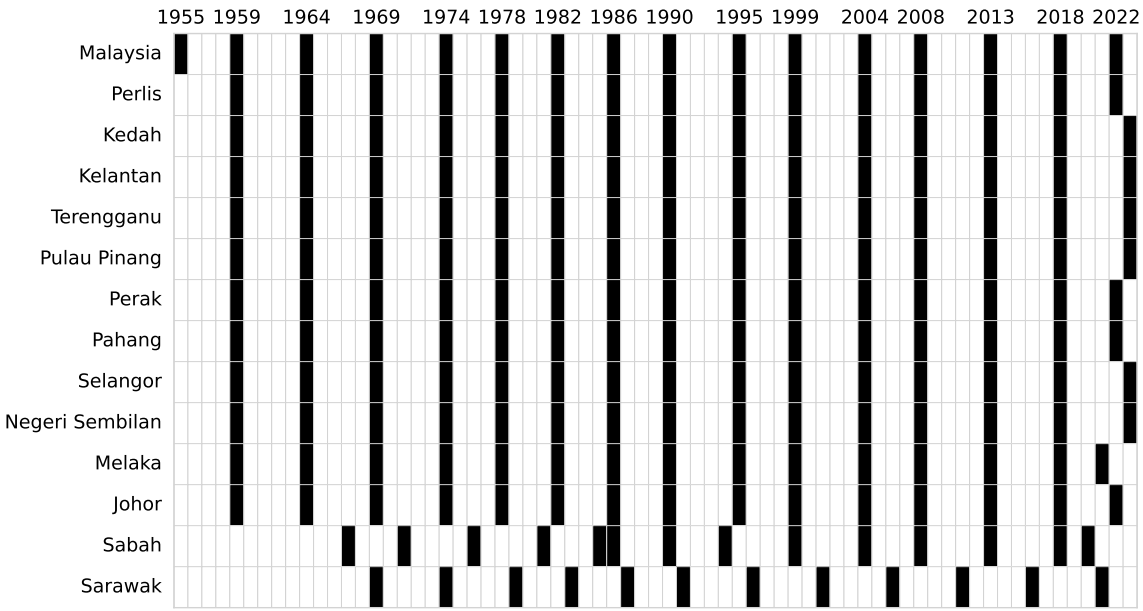
Empirical research and public knowledge on Malaysia's elections have long been constrained by a lack of high-quality open data, particularly in the absence of a Freedom of Information framework. This paper introduces the Malaysian Election Corpus (MECo), an open-access panel database covering all federal and state general elections since 1955, as well as by-elections since 2008. MECo includes candidate- and constituency-level data for 9,704 electoral contests across seven decades, standardised with unique identifiers for candidates, parties, and coalitions. The database also provides summary statistics for each contest (electorate size, voter turnout, majority size, rejected ballots, unreturned ballots), key demographic data for candidates (age, gender, ethnicity), and lineage data for political parties. MECo is the most well-curated open database on Malaysian elections to date, and will unlock new opportunities for research, data journalism, and civic engagement.

[1]Malaya University, W.P. Kuala Lumpur, Malaysia
✉ thevesh.theva@gmail.com

Background & Summary

Malaysia's electoral history is among the most dynamic in Southeast Asia, encompassing 2,715 federal election contests and 6,936 state election contests (Figure 1), as well as hundreds of off-cycle by-elections across a multiethnic, multi-party system. Furthermore, Malaysia offers significant scope for the study of democratisation, having experienced its first change of ruling party in 2018, and its first ever hung Parliament as recently as 2022. However, empirical studies of Malaysian elections are hindered by the lack of a comprehensive, standardised, and publicly available dataset that provides a single source of truth for scholars. The Election Commission (EC) does not publish open data which abides by best practices for data sharing, preferring instead to limit citizens to searching up isolated results via MySPR Semak. Gazetted election results published as subsidiary legislation by the Attorney General's Chambers (AGC) contain slightly more detail, but are available in PDF format only. The lack of a Freedom of Information Act further complicates efforts to acquire and systematically compile electoral returns.

Figure 1: Federal and state general election years



Amidst this paucity of data, global initiatives such as the Constituency-Level Elections Archive (CLEA)[1] provide immensely valuable cross-country coverage, including for Malaysian elections. However, they generally focus on federal contests, and within that scope, only on the number of votes won by each candidate (thus omitting information such as the electorate size, voter turnout, unreturned ballots, rejected ballots, etc). Similarly, international turnout or election integrity datasets[2,3] capture only high-level national indicators. Locally, a number of news and civil society organisations such as Tindak Malaysia,[4] malaysiakini (undi.info), and Sinar Project compile election data to varying degrees of completeness and quality, but these efforts—while laudable for their public service, and valuable as a stopgap measure—typically lack proper data hygiene and standardisation, and are not subjected to systematic validation and review, thus limiting their usefulness for rigorous empirical research and long-term preservation.

In this paper, I address this gap by providing the Malaysian Election Corpus (MECo),[5] the first complete open database of Malaysian election results at the federal and state level. MECo is intended as a living resource which provides the go-to empirical foundation for research and journalism on Malaysian elections. The database covers *all* general elections and by-elections since the pre-independence general election in 1955. In total, it records 14,000 unique candidates representing 110 political parties in 9,704 unique electoral contests from 1955 to the present.

The dataset comprises two core components:

- **Ballots**: The final results for each state legislative assembly constituency (DUN) and federal Parliament constituency (Parliament), with the number of votes received by each candidate.
- **Stats**: The electorate size, ballots issued, unreturned ballots, and ballots rejected in each constituency. For each constituency, I also derive the margin of victory (majority), voter turnout rate, ballot rejection rate, unreturned ballot rate, and majority as a share of valid votes.

Furthermore, my database offers three key advantages built on the use of unique identifiers (UIDs). First, I encode a UID for each candidate, allowing a single individual to be tracked across time even when they share a name with other candidates,

contest under different parties, or change the name used in public life; this is especially important in Malaysia, where politicians are not required to use their official legal names on electoral ballots. Second, I encode a UID for each party, enabling consistent tracking of political parties even when they undergo rebranding or organisational transformation, such as the evolution of the National Justice Party (PKN) into the People's Justice Party (PKR) in 2003 following a merger with the Malaysian People's Party (PRM). Third, I encode a UID for each coalition, allowing coalition membership to be identified separately from party identity; this is an essential distinction in Malaysia, where election ballots frequently list the coalition rather than the actual party of the candidate. In general, the use of UIDs makes the database highly extensible, allowing other researchers to build new lookup tables or enrich existing ones without needing to alter the core datasets.

To the best of my knowledge, no comparable database exists. As a living resource, this paper lays the foundation for future data curation, as well as research into areas like malapportionment, gerrymandering, local-level voting patterns, and the spatial dynamics of political competition. I also hope that MECo will serve as a catalyst for broader collaborations between academics, civil society, journalists, and election observers, supporting both scholarly inquiry and public accountability.

Finally, I note that while my work is the first of its kind for Malaysian elections, it follows a growing body of recent academic work focused on compiling and curating country-specific election data for reuse.[6–11] By situating MECo within this emerging tradition of high-quality electoral data curation, I contribute to the rapidly-improving global infrastructure of comparative political research. This work reflects a commitment to transparency, reproducibility, and the democratisation of access to electoral information.

Methods

There are three main data sources I used to construct my database:

1) Gazetted election results published in PDF format by the AGC; searching for "results of contested election" or "keputusan pilihan raya yang dipertandingkan" will yield the gazetted election results (Form 16, per

Regulation 27, Electoral (Conduct of Elections) Regulations 1981). The downloadable PDF contains Malay and English versions of all information.

2) Physical official election reports.[12–42]

3) MySPR Semak, an interactive website published by the EC; election results can only be queried one at a time, by selecting the appropriate election type (federal election, state election, or by-election), edition, and constituency. The site is only offered in Malay.

It should be noted that the first two sources are legally classified as open data and are not copyrightable under Malaysian copyright law. Section 3(1) of the Copyright Act 1987 (Act 332) expressly excludes from copyright "official texts of the Government or statutory bodies of a legislative or regulatory nature". The gazetted election results and official election reports fall under this category as formal statutory publications. In the case of the interactive dashboard, the dashboard itself is copyrighted by the EC, but the underlying data is entirely derived from the gazetted results, and is therefore open data. This interpretation was confirmed through consultation with legal advisors and officials familiar with election-document provenance. Accordingly, I archived PDFs of the gazetted election results and post-election reports I used to construct MECo (see Data Records).

I began with federal general elections, then state general elections, and finally by-elections. This is because state legislative assembly constituencies (DUNs) must lie completely within the boundaries of a federal parliamentary constituency (Parliament), so it was sensible to begin with the superset. By-elections coming last is an intuitive choice, since a by-election must follow a general election by definition; validation of by-election data is therefore dependent on having complete federal and state-level results.

Federal General Elections

For all federal elections in the dataset (Figure 2), I manually (i.e. by hand) digitised or copied the data from the aforementioned sources. I deliberately avoided the use of optical character recognition (OCR), PDF parsing, and web scraping tools after initial experimentation revealed an average error rate of approximately

Figure 2: Federal election coverage (number of seats)

	1955 (52)	1959 (104)	1964 (104)	1969 (144)	1974 (154)	1978 (154)	1982 (154)	1986 (177)	1990 (180)	1995 (192)	1999 (193)	2004 (219)	2008 (222)	2013 (222)	2018 (222)	2022 (222)
Sarawak	0	0	0	24	24	24	24	24	27	27	28	28	31	31	31	31
Johor	8	16	16	16	16	16	16	18	18	20	20	26	26	26	26	26
Sabah	0	0	0	16	16	16	16	20	20	20	20	25	25	25	25	25
Perak	10	20	20	20	21	21	21	23	23	23	23	24	24	24	24	24
Selangor	7	14	14	14	11	11	11	14	14	17	17	22	22	22	22	22
Kedah	6	12	12	12	13	13	13	14	14	15	15	15	15	15	15	15
Kelantan	5	10	10	10	12	12	12	13	13	14	14	14	14	14	14	14
Pahang	3	6	6	6	8	8	8	10	10	11	11	14	14	14	14	14
Pulau Pinang	4	8	8	8	9	9	9	11	11	11	11	13	13	13	13	13
W.P. Kuala Lumpur	0	0	0	0	5	5	5	7	7	10	10	11	11	11	11	11
Terengganu	3	6	6	6	7	7	7	8	8	8	8	8	8	8	8	8
Negeri Sembilan	3	6	6	6	6	6	6	7	7	7	7	8	8	8	8	8
Melaka	2	4	4	4	4	4	4	5	5	5	5	6	6	6	6	6
Perlis	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3
W.P. Putrajaya	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
W.P. Labuan	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1

10%, primarily due to frequent changes in formatting and layout, even within the same document. I considered this to be unacceptably high for a resource intended to serve as a single source of truth. Moreover, the downstream process of error detection and correction proved to be more time-consuming and error-prone than simply transcribing the data by hand, especially given the relatively manageable size of this data (25,545 rows of data across 9,704 electoral contests). The Technical Validation section further explains why the way in which Malaysia reports election results made it possible for me to do this with near-total confidence in the accuracy of the final data released for publication.

Records for seats in Peninsular Malaysia begin in 1955, while records for seats in Sabah and Sarawak begin in 1969. Although there was a federal general election in 1964, one year after Sabah and Sarawak (and Singapore) joined then-Malaya to form the Federation of Malaysia in 1963, seats in Sabah and Sarawak were not contested since the transition agreement allowed their respective state legislatures

to appoint (and not elect) their representatives to the federal Parliament.[43] There are no records for Singapore, which was not contested in the 1964 general election for the same reason as Sabah and Sarawak, and which exited the Federation prior to the next federal general election in 1969.

State General Elections

Figure 3: State election coverage (number of seats)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Sarawak	48	48	48	48	48	56	62	62	71	71	82	82			
Sabah	32	32	48	48	48	48	48	48	48	60	60	60	60	73	
Perak	40	40	40	42	42	42	46	46	52	52	59	59	59	59	59
Johor	32	32	32	32	32	32	36	36	40	40	56	56	56	56	56
Selangor	28	28	28	33	33	33	42	42	48	48	56	56	56	56	56
Kelantan	30	30	30	36	36	36	39	39	43	43	45	45	45	45	45
Pahang	24	24	24	32	32	32	33	33	38	38	42	42	42	42	42
Pulau Pinang	24	24	24	27	27	27	33	33	33	33	40	40	40	40	40
Kedah	24	24	24	26	26	26	28	28	36	36	36	36	36	36	36
Negeri Sembilan	24	24	24	24	24	24	28	28	32	32	36	36	36	36	36
Terengganu	24	24	24	28	28	28	32	32	32	32	32	32	32	32	32
Melaka	20	20	20	20	20	20	20	20	25	25	28	28	28	28	28
Perlis	12	12	12	12	12	12	14	14	15	15	15	15	15	15	15

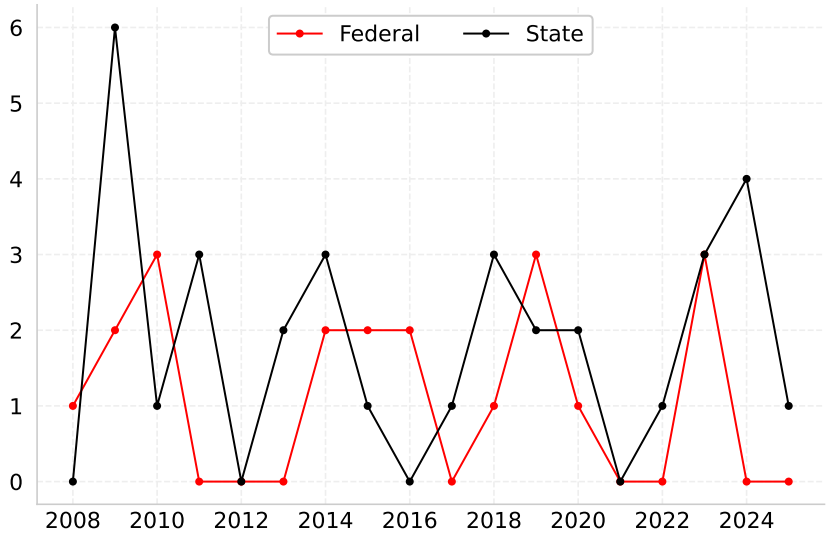
After completing the federal general elections dataset, I constructed the state general elections dataset in exactly the same way.

For states in Peninsular Malaysia, records begin in 1959, when general elections for all 13 state legislative assemblies were held concurrently with the federal general election. For Sabah and Sarawak, records begin in 1967 and 1969 respectively, the years of the first state general elections held after the formation of the Federation of Malaysia. In all, MECo contains records for 15 elections for all states in Peninsular Malaysia, 14 elections for Sabah, and 12 for Sarawak (Figure 3). The reason for the discrepancy between the number of observations for Sabah and Sarawak is that

there were two instances in Sabah's electoral history where state general elections were held in relatively quick succession. The first was in 1986, when Sabah went to the polls just one year after the previous state general election due to increasing civil and political instability.[44] The second was in 2020, when Sabah held a state general election two years after the watershed election of 2018 due to a collapse of the state government.

By-Elections

Figure 4: By-election coverage (number of elections)



Unlike for general elections, I could not locate any systematically compiled post-election reports or gazetted results for older by-elections. As a result, I relied solely on digital results made available via the EC's official website, which only covers by-elections since 2008. These digital records were copied, enriched, and standardised with the same schema used for general elections. Consequently, the by-elections dataset covers all Parliament and DUN by-elections held since the 12th federal general election in 2008, totaling 53 contests as of end-April 2025 (Figure 4).

I plan to expand the dataset to include an exhaustive record of previous by-elections once I am able to acquire a reliable source, likely via a combination of EC reports

and Hansards from Parliament and State Legislative Assemblies. However, given that the most recent three election cycles are already fully covered, the exclusion of older by-elections is likely negligible for almost all prospective users. I therefore made the decision not to delay the dissemination of MECo any further.

Unique Identifier (UID) and Lookup Table Generation

The database incorporates unique identifiers (UIDs) for parties, coalitions, and candidates to enable accurate longitudinal analysis. I also provide 3 lookup tables for parties, coalitions, and candidates, which allow users to augment the core datasets, either with variables I have already collated or with their own.

Parties and Coalitions. Before discussing the generation of party and coalition UIDs, it is critical to distinguish how 'party' is defined in this dataset relative to the EC's official records. When a candidate puts their name forth for nomination, they are mandated to declare themselves as either representing an entity registered with the EC, or as an independent candidate. However, the EC's list of registered parties does not distinguish between a single political party and a coalition of parties. For example, as of 30th October 2025, the list contains—without distinction—both the United Malays National Organisation (UMNO) as well as the Barisan Nasional (BN). The former is a 'party' in the conventional sense, while the latter is actually a coalition of several parties, including UMNO itself.

It is debatable as to whether the EC should amend their operating protocols to capture this distinction, or lower the barriers to a party being officially registered and greenlit for listing on a ballot. That judgment notwithstanding, the current set of practices has 3 particularly deleterious consequences for the quality of official election data:

1) **Misleading reporting of results for a single election**. Because the choice of party listed on the ballot is made at the candidate level, it is possible for candidates from the same party to list themselves differently. In practice, this happens because parties with a strong local identity choose to contest under the party flag within a particular state, but under the coalition flag in others. For example, the EC's official announcement of

the 2022 general election results listed PN and PAS as having won 52 and 22 seats respectively. In actuality, PAS was a component party of PN, but chose to contest in Kelantan and Terengganu under the party flag due to its deep roots in those states. This greatly increases the likelihood of erroneous conclusions by users of official data, especially if they lack the required contextual knowledge.

2) **Loss of insight into party dynamics**. Political parties which are part of stable coalitions often choose to contest under the coalitional identity. For example, UMNO—Malaysia's oldest political party—has never contested an election under its own banner, instead contesting under the Alliance flag until 1973, and the BN flag thereafter. This renders it impossible to analyse the political trajectory of UMNO relative to other component parties (especially MIC and MCA) using official data.

3) **Loss of insight into unofficial coalitions**. For example, Pakatan Harapan was not officially registered in time for the 2018 general election, thus leading to candidates from its 3 component parties (PKR, DAP, and AMANAH) mostly choosing to contest under the PKR flag. While it is true that contemporaneous election observers were generally fully aware of this circumstance, future researchers may come to a different (incorrect) conclusion if they rely solely on official records.

For this reason, I made the methodological choice to capture 3 separate variables (see Data Records for complete schema). First, `party_on_ballot`—which faithfully captures the exact listing of the candidate on the official ballot. Second, `party`—which captures the true party allegiance of the candidate. Third, `coalition`—which captures the coalitional allegiance of the party. It was extremely difficult to consolidate this information going back to 1955, especially for state-level results which receive far less attention than federal results. However, I was able to eventually fill in all gaps using a combination of existing scholarly work,[45–48] media reports (for modern elections), archived news articles (for older elections), and consultation with colleagues specialising in election history. To the best of my knowledge, MECo is now the only publicly-available dataset with complete data allowing users to distinguish between the 3 dimensions of political allegiance captured here.

Once this was done, party and coalition UIDs were very easy to generate and validate manually, given that the total number of distinct parties (110) and coalitions (18) in the dataset is relatively small. The UIDs for coalitions are simple integers, while the UIDs for parties take the following form (e.g. `001-UMNO`):

`{integer}-{acronym}`

The reason I implemented this syntax is to encode sufficient information for 'versioning' within the UID, given that instances of renaming and rebranding are fairly common in Malaysia's political landscape. For example, the Federated Sabah People's Front (UID: `065-BERSEKUTU`) was founded in 1994, renamed to the Sabah People's Front (`065-SPF`) in 2010, taken over and rebranded to the Sarawak Workers Party (`065-SWP`) in 2012, and then renamed again to the present-day Malaysian Nation Party (`065-PBM`) in 2021. An integer-only approach would have necessitated either 'collapsing' these 4 iterations into one UID (resulting in loss of information), or using four different UIDs (resulting in loss of ability to chain versions). My solution enables users to detect that these 4 iterations are part of the same chain (via the 065 prefix), while still maintaining 4 separate rows in a lookup table so that information on each iteration can be provided. This syntax is also handy for distinguishing between parties using the same acronym; for instance, 'UPKO' maps to 3 distinct parties which existed at different points in Sabah's election history.

Finally, for political parties, I used the UIDs as the basis for creating a separate lookup table tracking the lineage of political parties, i.e. noting down instances of merging, splitting, splintering, and renaming/rebranding linked to the UIDs of the predecessor and successor respectively. I did not deem this necessary for coalitions, which should be analysed by examining their party composition during elections, since coalition membership is always in flux and is independent of changes in the core identity of the coalition.

Candidates. Candidate UIDs enable the accurate tracking of individual candidates across elections. Beyond the obvious use case of distinguishing candidates with the same name (e.g. the dataset contains 8 unique individuals with the exact same name—'Ahmad bin Abdullah'), robust candidate UIDs are particularly important in the Malaysian context, where candidates frequently acquire new honorifics, adopt

different formatting conventions, or spell their name differently over time. Two illustrative examples demonstrate the need for candidate UIDs. 8-time Parliamentarian **Rafidah Aziz** appeared on ballots across 8 elections in forms including:

Rafidah Aziz
Rafidah Bt. Ab. Aziz
Rafidah Bt. Abdul Aziz
Datuk Seri Rafidah Aziz

Another 8-time Parliamentarian **Samy Vellu A/L Sangalimuthu** appeared on ballots across 9 elections with variations such as:

S. Samy Vellu
Datuk Seri Samy Vellu
S. Samy Vellu A/L Sangalimuthu
Dato' S. Samy Vellu A/L Sangalimuthu

Through careful examination, I assigned a single UID to all instances of each candidate, taking particular care to capture all permutations of an individual's name while distinguishing between different individuals who happened to share a name. Two cues were especially useful: *space* and *time*. Consecutive elections in the same seat often revealed consistent candidacy patterns (i.e. a 'home base'), whereas large gaps across states or decades typically indicated distinct individuals despite identical names. These cues guided targeted searches of external information to ensure that each UID was assigned correctly. Prominent candidates were straightforward to verify, as anyone who won an election appears in public records. The most challenging cases were independent candidates who contested only once, especially in older elections prior to widespread digital documentation. The full process of assigning candidate UIDs took nearly a year; in MECo's present form, I am confident that most, if not all, detectable errors have been eliminated.

Once a complete and validated set of candidate UIDs was established, I created a candidate lookup mapping each UID to a cleaned name stripped of titles and honorifics. This lookup was then enriched with three demographic attributes: sex, ethnicity, and date of birth (DOB). Sex and ethnicity were assigned via manual inspection of names, which are nearly perfectly indicative of sex and strongly indicative

of ethnic group in the Malaysian context; both variables are complete for 100% of candidates. Where possible, these were verified against public records, particularly for recent elections. DOB were obtained through extensive manual searches across thousands of sources—including biographical directories, news archives, parliamentary profiles, and obituary notices—and recorded in ISO format (YYYY-MM-DD). Reliable birth information could not be found for 33.2% of candidates overall, reflecting limited public documentation in earlier decades. Coverage improves markedly over time: 60% of candidates in GE3 (1969) lack DOB information, falling to 16% by GE10 (1999) and under 5% by GE15 (2022), with remaining gaps concentrated among independent candidates with minimal public profiles.

No UID for Constituencies

Constituency names are not reliable indicators of spatial continuity. The name of a constituency may remain the same despite substantial spatial change, as in the case of Lumut (Perak) in the 2018 delimitation exercise. Similarly, a constituency may be renamed despite the underlying territory remaining largely intact, as in the case of Silam (Sabah) being renamed to Lahad Datu in the 2019 delimitation exercise. More fundamentally, as several streams of work in the field have illustrated,[49–51] the notion of what makes a constituency the same or different across elections is inherently subjective, and should be calibrated to the specific research question at hand. Therefore, any attempt at generating a constituency UID using names alone would be a rough attempt at best, and incredibly misleading at worst.

Consequently, I consider constituency continuity best left to researchers. In fact, it arguably merits dedicated scholarly handling, since electoral lineage has never been rigorously documented or studied in the Malaysian context. That having been said, users who wish to augment MECo with other constituency-based datasets—for example, the subnational statistics published by the Department of Statistics Malaysia—should rely on the intrinsic uniqueness of the `date-state-constituency` combination as a composite key. This avoids imposing any particular lineage model, while ensuring that external data can be merged cleanly and reproducibly.

Data Records

All datasets (Table 1) are published on Harvard Dataverse,[5] which serves as the canonical archive. For convenience, the exact same datasets are also mirrored on:

- **Zenodo**[52] (doi.org/10.5281/zenodo.17694675)
  Provides code and raw source files, in addition to datasets.
- **GitHub** (github.com/Thevesh/paper-malaysian-election-corpus)
  Facilites active development and maintenance, issue tracking and community contributions. Substantial updates are released on Zenodo.

The datasets fall into two groups; constituency-level statistics and candidate-level ballots, followed by lookup tables that extend the core schema.

Table 1: Description of primary datasets

Filename	**Description**
`consol_stats`	Summary statistics for all federal and state elections
`federal_stats`	Subset of `consol_stats` for federal general elections
`state_***_stats`	Subset of `consol_stats` for state general elections
`byeelection_stats`	Subset of `consol_stats` for by-elections
`consol_ballots`	Candidate-level results for all federal and state elections
`federal_ballots`	Subset of `consol_ballots` for federal general elections
`state_***_ballots`	Subset of `consol_ballots` for state general elections
`byeelection_ballots`	Subset of `consol_ballots` for by-elections
`lookup_candidate`	Standardised list of candidates
`lookup_party`	Standardised list of parties
`lookup_party_succession`	Details of merging, splitting, splintering and rebranding
`lookup_coalition`	Standardised list of coalitions
`lookup_dates`	Election dates, by state and election type
`logs/corrections`	Log of manual corrections applied to `ballots_issued` as described in the Technical Validation section

Table 2 provides a detailed description of the `*_stats` files. Each row corresponds to a single constituency in a single election, and contains the complete set of numerical aggregates for that contest. These files are therefore the best entry point for constituency-level analysis or for merging with external datasets organised at the constituency level. Users should note that the `*_stats` files have a one-to-many relationship with the corresponding `*_ballots` files; each row in a `*_stats` file links to multiple candidate-level rows in the `*_ballots` file, except in uncontested seats where only one candidate appears. The two file types can be joined using the composite key `(date, state, seat)`.

Table 2: Structure of all `*_stats` files

Variable	**Description**
`date`	Date of the election (YYYY-MM-DD)
`election`	Election name (e.g., GE-14, SE-10, BY-ELECTION)
`state`	State in which the seat is located
`seat`	Full name of the seat (e.g., P.049 Tanjong)
`voters_total`	Total number of registered voters
`ballots_issued`	Number of ballots issued
`ballots_not_returned`	Number of ballots not returned (often postal votes)
`votes_rejected`	Number of rejected (spoiled) ballots
`votes_valid`	Number of valid votes
`majority`	Margin of victory (winner minus runner-up)
`n_candidates`	Number of candidates contesting
`voter_turnout`	Ballots issued as a share of registered voters (%)
`majority_perc`	Majority as a share of valid votes (%)
`votes_rejected_perc`	Rejected ballots as a share of ballots returned (%)
`ballots_not_returned_perc`	Unreturned ballots as a share of ballots issued (%)

Table 3 provides a detailed description of the `*_ballots` files. Each row corresponds to a single candidate contesting a specific constituency in a specific election. These files are therefore the best entry point for candidate-level analysis or for merging with external datasets organised at the candidate level. Party- or coalition-level analysis should also begin from these files.

Table 3: Structure of all `*_ballots` files

Variable	**Description**
`date`	Date of the election (YYYY-MM-DD)
`election`	Election name (e.g., GE-14, SE-10, BY-ELECTION)
`state`	State in which the constituency is located
`seat`	Code and full name of the seat (e.g., P.052 Bayan Baru)
`ballot_order`	Order in which the candidate appeared on the ballot
`candidate_uid`	Unique identifier for the candidate
`name_on_ballot`	Candidate name as it appeared on the ballot paper
`party_on_ballot`	Party name as it appeared on the ballot paper
`party_uid`	Unique identifier for the party
`party`	True party allegiance of the candidate
`coalition_uid`	Unique identifier for the coalition
`coalition`	True coalitional allegiance of the party
`votes`	Number of valid votes received by the candidate
`votes_perc`	Share of valid votes received by the candidate (%)
`rank`	Rank of the candidate in that contest
`result`	Outcome (won, won uncontested, lost, lost deposit)

In addition to the primary ballots and statistics files, the database includes 5 lookup tables that support the core datasets. These lookup tables provide a structured way

for users to perform deeper analysis or integrate external datasets without needing to modify the core data, thus guarding against downstream errors.

- `lookup_candidate`: Standardised list of all candidates, including cleaned names (stripped of titles and honorifics) and demographic attributes (sex, ethnicity, and date of birth). This table should be left-joined onto any `*_ballots` file using `candidate_uid` as a key.
- `lookup_party`: Standardised list of political parties, including party acronyms, full party names, and year of formation. This table should be left-joined onto any `*_ballots` file using `party_uid` as a key.
- `lookup_party_succession`: A lineage table documenting instances of parties merging, splitting, splintering, being absorbed, or rebranding. The `predecessor_uid` and `successor_uid` fields are fully consistent with the `party_uid` field in the `lookup_party` table, thus enabling joining in either direction as required.
- `lookup_coalition`: Standardised list of electoral coalitions, including coalition UIDs, names, and short descriptions. This table should be left-joined onto any `*_ballots` file using `coalition_uid` as a key.
- `lookup_dates`: A complete listing of all election dates for federal, state, and by-election contests. This table is provided to facilitate preparation of temporal joins with external datasets.

Technical Validation

There are five critical components of the database which require validation: Numerical data, candidate names, parties, coalitions, and constituencies.

For numerical data, the way in which Malaysia reports election results makes a

unique form of validation possible. First, I define the following variables:

$$\begin{aligned} I &= \text{Ballots Issued} \\ U &= \text{Unreturned Ballots} \\ R &= \text{Ballots Rejected} \\ V_i &= \text{Valid Votes for Candidate } i \end{aligned}$$

For a given contest involving N candidates, the following relationship must hold:

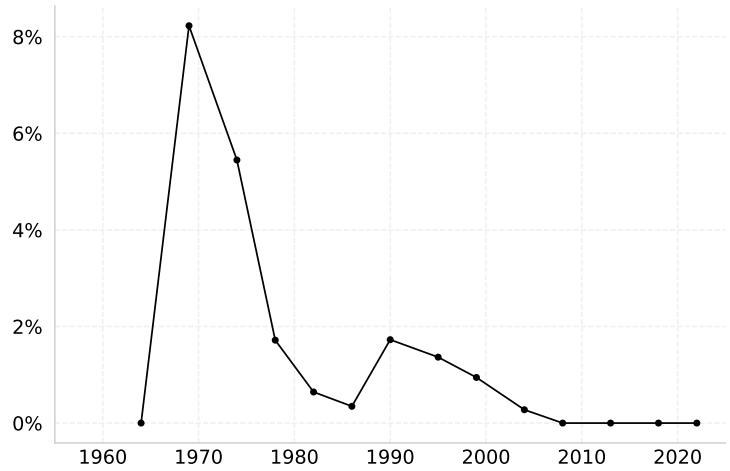
$$I - U - R = \sum_{i=1}^{N} V_i \tag{1}$$

In plain language, the implied number of valid votes must be equal to the sum of votes received by all candidates. Because I, U, R, and V_i are (and have been) reported separately in historical election reports and gazetted results, I can leverage this relationship as an almost-foolproof way to validate the accuracy of my manual data entry process. For transparency, I detected and corrected errors for 82 out of 9,704 contests, implying an error rate of 0.84%, which is significantly lower than the 10% error rate I encountered in my initial OCR experiments. Importantly, these errors are detectable and fixable, and come without the additional overhead of correcting errors in text data; my familiarity with local geography, politicians, and culture enabled me to transcribe text data rapidly and accurately.

This validation procedure also revealed 114 contests for which the data was digitised accurately based on my source material, but which nevertheless failed to satisfy Equation (1). These errors presumably occurred due to mistakes in data entry which were not caught and corrected during the original publication process. In order to ensure a clean dataset, I applied a standard correction—for all 114 contests failing validation, I adjusted the number of ballots issued (as documented in `corrections.csv`) such that Equation (1) holds. I chose `ballots_issued` as the variable to adjust for two reasons. First, until the 1981 amendment of the Elections (Conduct of Elections) Regulations (implemented fully from the 1990 federal general election onwards), the number of unreturned ballots was not reported (and thus assumed to be 0). Second, because the number of rejected ballots is very small relative to the number of ballots issued, using rejected ballots as the adjusted variable would have resulted in a much larger correction in percentage terms—specifically,

the correction would have required a 20% change to the number of rejected ballots on average (with several exceeding 50%), relative to a 0.6% change to the number of ballots issued (with only one above 5%, and none above 20%).

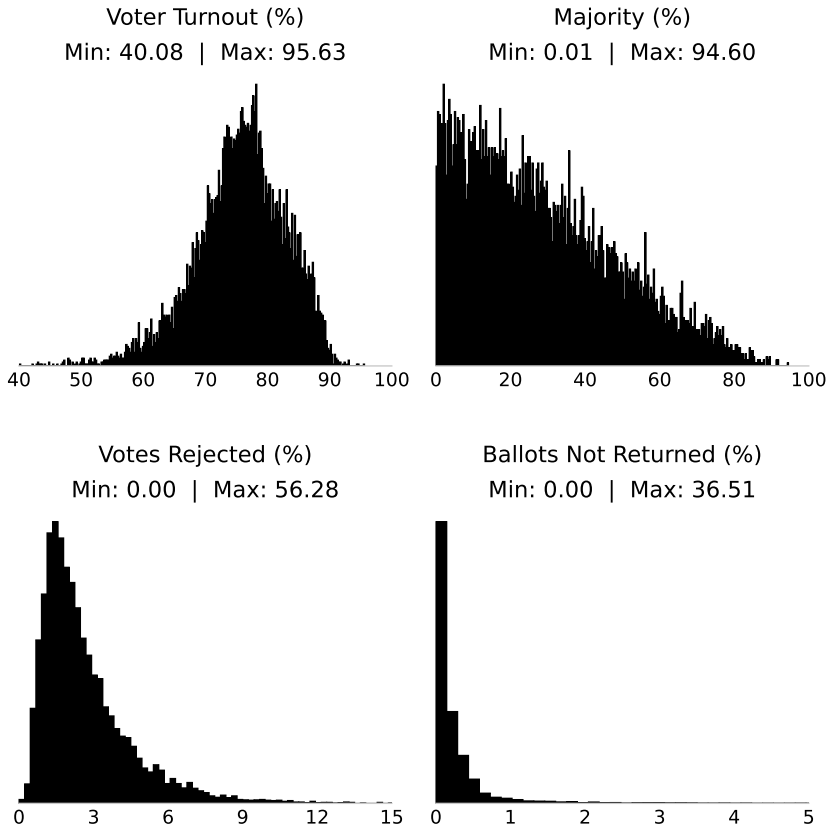
Figure 5: Error rate in general election years



An interesting trend emerges when I plot the error rate arising from these 114 contests (Figure 5). The error rate was only 0.26% in 1964 (just 1 error), but spiked to nearly 8% in 1969, when elections were severely disrupted by the Emergency. By 1978, error rates stabilised below 2%, and have been constant at 0% since 2008. I posit that this reflects general improvements in data management technology over the decades; I do not have a clear explanation for the absence of errors in 1964.

Wrapping up my checks on numerical values, I note that the validation procedure can fail if two errors exactly offset each other. Furthermore, it cannot detect errors in the number of registered voters, which does not enter Equation (1). To address these limitations, I plotted histograms of 4 derived variables (Figure 6): voter turnout rate, ballot rejection rate, unreturned ballot rate, and majority as a share of valid votes. All four variables pass the check of being bounded between 0 and 100%, and display smooth distributions as would be expected if data entry was accurate. All outlier values were double-checked; in particular, I verified that extreme instances of high rejected ballots or unreturned ballots were not due to mistakes in data entry. For example, the DUN of Pangkor in Perak had two instances where over 30% of ballots

Figure 6: Histogram of derived variables



were not returned; this was due to historical inefficiencies in the implementation of postal voting for navy personnel.[53]

For non-numerical data, I conducted 3 types of checks. For parties and coalitions, the full dataset only contains 110 parties and 18 coalitions, so I manually checked each against publicly available information. For constituency names, I generated a list of unique constituency names (1,437 in total), which I manually checked for errors in spelling or syntax. The most challenging component was candidate names and demographic details due to data volume (25,545 rows). Although I applied rigorous formatting standards during data entry and conducted a full round of manual checking (line-by-line), I acknowledge that minor inconsistencies or inaccuracies may persist. This is especially because candidates often change the presentation of their names across elections (e.g., spelling out vs shortening a surname), acquire new ti-

tles and honorifics over time (e.g., Dato', Tan Sri, Haji, Ustaz, academic degrees), and format their names differently in different years (e.g., omitting or reordering titles). Such variations make consistent longitudinal tracking inherently complex. As with all living datasets of this scale, I anticipate incremental improvements over time as users engage with the data and identify potential refinements.

Finally, to ensure accuracy of the database as a whole, I derive seats and votes by party for all elections in the dataset, and ensure that these match against the record of parties in Parliament and State Legislative Assemblies.

Usage Notes

The Malaysian Election Corpus (MECo) is designed to support a wide range of use cases, ranging from rigorous empirical research to rapid data journalism, and even casual civic technology projects. In the academic realm, MECo —as the first database of its kind for Malaysia—serves as a foundational resource for research in electoral studies, political science, and public policy. Researchers can employ the data to answer important questions about the evolution of Malaysia's electoral system, employing both cross-sectional and longitudinal analysis.

The data can also be interactively explored via ElectionData.MY. In addition to making the data accessible to non-technical users, the site is intended to serve as the primary channel for continuous updates and enhancements. Future improvements to MECo will be reflected on the site as new versions are archived on Harvard Dataverse[5] and Zenodo,[52] ensuring that users always have the freshest data.

I anticipate that many users will want to extend or adapt the dataset. The standardised schema and use of lookup tables enable seamless enrichment and integration with other datasets. Some valuable examples include:

- `lookup_candidate` can be enriched with additional demographic or biographical information such as marital status, education, and occupation.
- `lookup_party` and `lookup_coalition` can be extended with variables such as ideological classification, membership size, or even beneficial ownership.

- Constituency-level data can be merged directly with MECo using the intrinsic uniqueness of the `date-state-constituency` combination, thus enabling the use of MECo in conjunction with official or alternative datasets.

As a practical reference for users, the codebase contains samples (`dashboards.py`) of how to merge the core datasets with lookup tables to create panel data suitable for interactive dashboards and advanced analyses.

Finally, while rigorous validation has been applied (see Technical Validation), this remains a living database and is intended as such. Minor inaccuracies, particularly in the candidate name field, may persist as discussed above. I encourage users to report issues or submit improvements via the GitHub repository, which provides full transparency of changes between releases.

Data Availability

All datasets described in this paper are published on Harvard Dataverse[5] under a CC0 license. For convenience, the exact same datasets are also mirrored in repositories on Zenodo[52] and GitHub. Both contain code and raw source files in addition to the datasets; I use GitHub to version-control active development and maintenance, with substantial updates released on Zenodo.

Code Availability

All data processing, validation, and compilation was conducted in Python. The full source code is publicly available under a CC0 license via Zenodo[52] and GitHub. Three key scripts which users should find particularly useful are:

- `compile.py`, which generates and validates all tabular datasets, with the exception of the `lookup*.csv` files, which were manually curated.
- `dataviz.py`, which generates all visualisations used in this paper.
- `dashboards.py`, which generates the source files for the interactive visualisations available at ElectionData.MY.

References

[1] Ken Kollman, Allen Hicken, Daniele Caramani, David Backer, and David Lublin. Constituency-Level Elections Archive (CLEA). https://doi.org/10.17616/R33S92, 2024.

[2] Ferran Martínez i Coma and Diego Leiva Van De Maele. The global dataset on turnout (GD-Turnout). *Electoral Studies*, 86:102681, 2023.

[3] Pippa Norris, Richard W Frank, and Ferran Martínez i Coma. Measuring electoral integrity around the world: A new dataset. *PS: Political Science & Politics*, 47(4):789–798, 2014.

[4] Danesh Prakash Chacko. Tindak Malaysia Historical Election Results. https://doi.org/10.5281/zenodo.17693558, 2025.

[5] Thevesh Thevananthan. The Malaysian Election Corpus: Federal and state-level election results since 1955, v5. https://doi.org/10.7910/DVN/O4CRXK, 2025.

[6] Virgilio Pérez, Cristina Aybar, and Jose M Pavía. Spanish electoral archive. SEA database. *Scientific Data*, 8(1):193, 2021.

[7] Samuel Baltz, Alexander Agadjanian, Declan Chin, John Curiel, Kevin DeLuca, James Dunham, Jennifer Miranda, Connor Halloran Phillips, Annabel Uhlman, Cameron Wimpy, et al. American election results at the precinct level. *Scientific Data*, 9(1):651, 2022.

[8] Justin de Benedictis-Kessner, Diana Da In Lee, Yamil R Velez, and Christopher Warshaw. American local government elections database. *Scientific Data*, 10(1):912, 2023.

[9] Bruno Calderón-Hernández, Horacio Larreguy, John Marshall, and José Luis Pérez-Castellanos. Electoral precinct-level database for Mexican municipal elections. *Scientific Data*, 12(1):582, 2025.

[10] Vincent Heddesheimer, Hanno Hilbig, Florian Sichart, and Andreas Wiedemann. GERDA: The German election database. *Scientific Data*, 12(1):618, 2025.

[11] Francesca R Jensenius and Gilles Verniers. Studying Indian politics with large-scale data: Indian election data 1961–today. *Studies in Indian Politics*, 5(2):269–275, 2017.

[12] T. E. Smith. *Report on the First Election of Members to the Legislative Council of the Federation of Malaya*. Supervisor of Elections, 1955.

[13] Election Commission. *Report on the Parliamentary and State Elections 1959*. Election Commission of the Federation of Malaya, 1959.

[14] Election Commission. *Report on the Parliamentary (Dewan Ra'ayat) and State Legislative Assembly General Elections, 1964 of the States of Malaya*. Election Commission of the Federation of Malaya, 1964.

[15] Election Commission. *Report on the Parliamentary (Dewan Ra'ayat) and State Legislative Assembly General Elections 1969 of the States of Malaya, Sabah and Sarawak*. Election Commission of Malaysia, 1971.

[16] Election Commission. *Report on the Parliamentary (Dewan Rakyat) and State Legislative Assembly General Elections 1974 of the States of Malaya and Sarawak*. Election Commission of Malaysia, 1974.

[17] Election Commission. *Report on the General Elections to the House of Representatives and the State Legislative Assemblies Other Than the State Legislative Assemblies of Kelantan, Sabah and Sarawak 1978*. Election Commission of Malaysia, 1978.

[18] Election Commission. *Report on the Malaysian General Elections 1982*. Election Commission of Malaysia, 1982.

[19] Election Commission. *Report on the Malaysian General Elections 1986*. Election Commission of Malaysia, 1986.

[20] Election Commission. *Report on the Malaysian General Elections 1990*. Election Commission of Malaysia, 1990.

[21] Election Commission. *Report of the General Election Malaysia 1995*. Election Commission of Malaysia, 1995.

[22] Election Commission. *Report of the General Election Malaysia 1999*. Election Commission of Malaysia, 1999.

[23] Election Commission. *Report of the General Election Malaysia 2004*. Election Commission of Malaysia, 2004.

[24] Election Commission. *Report of the General Election Malaysia 2008*. Election Commission of Malaysia, 2008.

[25] Election Commission. *Report of the General Election Malaysia 2013*. Election Commission of Malaysia, 2013.

[26] Election Commission. *Report on the State Legislative Assembly General Election 1967 of the State of Sabah*. Election Commission of Malaysia, 1967.

[27] Election Commission. *Report on the State Legislative Assembly General Election 1971 of the State of Sabah*. Election Commission of Malaysia, 1971.

[28] Election Commission. *Report on the State Legislative Assembly General Election 1976 of the State of Sabah*. Election Commission of Malaysia, 1976.

[29] Election Commission. *Report on the State Legislative Assembly General Election 1981 of the State of Sabah*. Election Commission of Malaysia, 1981.

[30] Election Commission. *Report on the State Legislative Assembly General Election 1985 of the State of Sabah*. Election Commission of Malaysia, 1985.

[31] Election Commission. *Report on the State Legislative Assembly General Election 1986 of the State of Sabah*. Election Commission of Malaysia, 1986.

[32] Election Commission. *Report on the State Legislative Assembly General Election 1990 of the State of Sabah*. Election Commission of Malaysia, 1990.

[33] Election Commission. *Report on the State Legislative Assembly General Election 1994 of the State of Sabah*. Election Commission of Malaysia, 1994.

[34] Election Commission. *Report on the State Legislative Assembly General Election 1999 of the State of Sabah*. Election Commission of Malaysia, 1999.

[35] Election Commission. *Report on the State Legislative Assembly General Election 1979 of the State of Sarawak*. Election Commission of Malaysia, 1979.

[36] Election Commission. *Report on the State Legislative Assembly General Election 1983 of the State of Sarawak*. Election Commission of Malaysia, 1983.

[37] Election Commission. *Report on the State Legislative Assembly General Election 1987 of the State of Sarawak*. Election Commission of Malaysia, 1987.

[38] Election Commission. *Report on the State Legislative Assembly General Election 1991 of the State of Sarawak*. Election Commission of Malaysia, 1991.

[39] Election Commission. *Report on the State Legislative Assembly General Election 1996 of the State of Sarawak*. Election Commission of Malaysia, 1996.

[40] Election Commission. *Report on the State Legislative Assembly General Election 2001 of the State of Sarawak*. Election Commission of Malaysia, 2001.

[41] Election Commission. *Report on the State Legislative Assembly General Election 2006 of the State of Sarawak*. Election Commission of Malaysia, 2006.

[42] Election Commission. *Report on the State Legislative Assembly General Election 1978 of the State of Kelantan*. Election Commission of Malaysia, 1978.

[43] Gordon P. Means. *Malaysia: The Second Generation*. Oxford University Press, Singapore, 1991.

[44] Mohammad Agus Yusoff. Sabah politics under Pairin. *JATI - Journal of Southeast Asian Studies*, 6:29–48, 2001.

[45] Bridget Welsh. *Elections in Malaysia*. Routledge Handbook of Contemporary Malaysia, 2016.

[46] Lim Hong Hai and Ong Kian Ming. Electoral campaigning in malaysia. In *Election Campaigning in East and Southeast Asia*, pages 55–77. Routledge, 2017.

[47] Francis Kok-Wah Loh and Boo Teik Khoo. *Democracy in Malaysia: discourses and practices*. Number 5. Psychology Press, 2002.

[48] Dieter Nohlen, Florian Grotz, and Christof Hartmann. *Elections in Asia and the Pacific: A data handbook*, volume 2. Oxford University Press, 2001.

[49] Ursula Hackett. Redrawing democracy: Quantifying house district continuity and change, 1789–2024. *Studies in American Political Development*, pages 1–15, 2025.

[50] Michael H Crespin. Using geographic information systems to measure district change, 2000–2002. *Political Analysis*, 13(3):253–260, 2005.

[51] Eric McGhee. Measuring efficiency in redistricting. *Election Law Journal: Rules, Politics, and Policy*, 16(4):417–442, 2017.

[52] Thevesh Thevananthan. Thevesh/paper-meco-results. https://doi.org/10.5281/zenodo.17694675, 2025.

[53] Iskandar Dzulkarnain Ahmad Junid. The battle for Perak. In *Malaysia's 14th General Election and UMNO's Fall*, pages 223–239. Routledge, 2019.

Acknowledgements

I thank Rosmadi Fauzi and Zulkanain Abdul Rahman for their helpful feedback and suggestions on the dataset and manuscript. I also gratefully acknowledge Nimesha Thevananthan for her assistance with data collection, Danesh Prakash Chacko of Tindak Malaysia for providing historical insights and filling data gaps, and the MyElectionData volunteers for their assistance with completion and validation of candidate UID harmonisation and candidate demographic data. Finally, I thank the EC for granting access to physical election reports via *Pusat Sumber SPR*.

Author Contributions (CRediT statement)

TT: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review and editing, Visualization, Supervision, Project administration.

Funding

This work did not receive any specific funding.

Competing Interests

The author declares no competing interests. The author is not affiliated with any political party or election-based organisation.