Chemical Risk Anticipation Tool Validation

Outcome 2: Automating the extraction of chemical prevalences from a bibliographic database to estimate "emerging concern" — a pilot study

Jason M Whyte[1,2] and Andrew P Robinson[1]

[1]Centre of Excellence for Biosecurity Risk Analysis, The University of Melbourne
[2]ACEMS at The School of Mathematics and Statistics, The University of Melbourne

January 30, 2020

cebra
Centre of Excellence for Biosecurity Risk Analysis

Contents

List of Figures

List of Tables

1. Executive summary

The New South Wales Environment Protection Authority (NSW EPA) commissioned the development of a Chemical Prioritisation Framework (CPF) in order to focus greater regulatory attention on potential risks to human and animal welfare. The framework requires the "emerging concern" associated with candidate chemicals as an input. The emerging concern associated with a given chemical was derived from the number of publications in the Taylor and Francis online portal referring to that chemical over a calendar year ("hits"), for a period of years. This limited the search for hits to those by only one publisher, and hit counts were drawn manually.

Following a review of the CPF methodology (Outcome 1 of the current study), the Centre of Excellence for Biosecurity Risk Analysis (CEBRA) proposed some extensions and refinements. We investigated these in a systematic manner that is reproducible and largely automated via this pilot study. We achieved this by using R code to obtain hit counts over calendar years from the Web of Science core collection API (henceforth, WOS).

This study conducted a successful initial exploration of the "technological space" of data sources, and means of extracting information from them. Our methodology can:

- recognise certain trends in scientific publication data, and,
- discern a point in time at which there is a sharp increase in the attention given by the scientific literature to the harmful effects of a chemical — the type of event one could associate with "emerging concern".

We asked NSW EPA to provide a list of chemicals of interest, and to classify these as being of anticipated low (under 100 hits), medium (from 100 up to 1000 hits), and high (from 1000 up to 10 000 hits) prevalence for the period from 2013 to 2017. From this list we randomly selected six chemicals (two from each of the three prevalence groups) and investigated their prevalence in WOS. Using a "Topic Search" allowed us to search for hits in certain fields of records, including titles, abstracts, and lists of keywords. We considered hit counts by year of publication over suggested year ranges. Briefly, our findings for the tasks agreed for this project are as follows:

1. Time series of counts of prevalences to inform the modelling of future concern:
   We queried WOS to determine the "total-hits" (any mention) for our selected chemicals from 2012 to 2018. We expected that this may include publications unrelated to some adverse effect. As such, we sought to explore alternatives by requiring that a hit must have a chemical name appearing in the same field as ("same-hits"), or within three words of ("near-hits") search modifiers relating to harmfulness. Finally, to enable comparison against the previous NSW EPA results, we queried WOS to find those hits that had a digital object identifier beginning with the prefix used by Taylor and Francis publications ("total-T&F-hits").
   Hit counts (shown in Figure 4.1 of the report) are reproduced below in Figure 1.1.

In some cases there is a substantial difference between the hit counts obtained for the different types of WOS queries applied for a given chemical. For example, for 2,3,7,8-tetrachlorodibenzo-p-dioxin the total-T&F-hits are substantially lower than those for near-hits. Such cases are notable as — even though the query for near-hits is quite restrictive — they show that WOS has greater coverage than the Taylor & Francis database, and thus WOS is more suitable for this application.

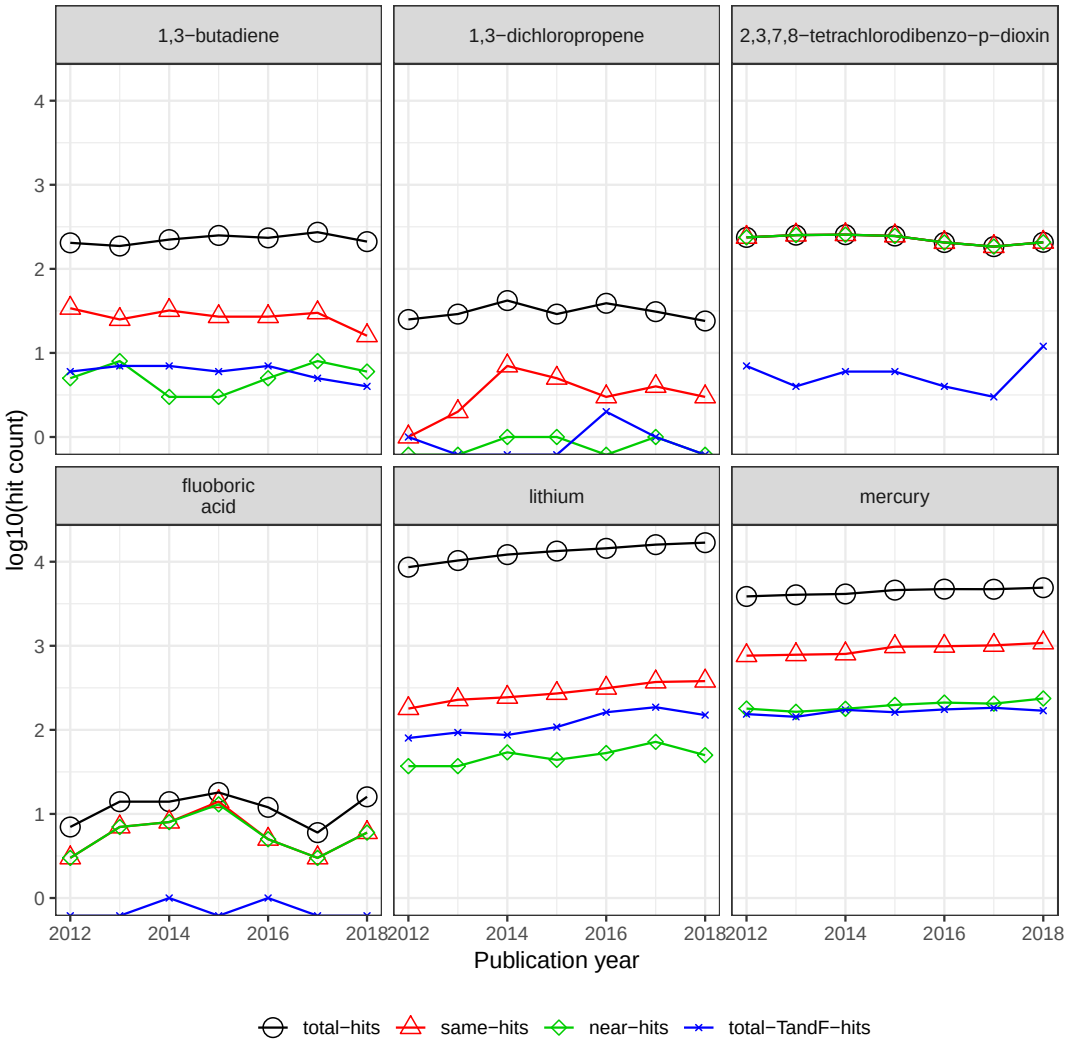

Figure 1.1.: Time series of the four hit types in WOS for the sample of chemicals from 2012-2018. Whenever fewer than four series are shown on a panel, some series are overlaid exactly by the "near-hits" series.

We experimented with some modelling of time series of near-hits counts using simple time series models (Figure 5.1). For five of six chemicals, we can reasonably predict the 2018 near-hit count using a time series model fit to the 2012-2016 data. We can use our models to judge whether a chemical should be designated as being of *low*, *moderate*, or *high* prevalence in 2018, and comparison against data has validated our process. We see this as a positive indication that our approach can develop into a means by which to judge which chemicals are appropriate priorities for monitoring.

2. Time series of counts of prevalences in detecting when historical publication data

relating to a chemical demonstrates a rapid growth in concern:
Continuing our use of WOS, we compared a measure of emerging concern similar to that previously employed by NSW EPA with near-hits for our six sample chemicals for the publication range of 1988 to 2018. Comparing the time series of these measures for a given chemical shows a notable difference in their behaviour (Figures 6.1 and 6.2). We believe that near-hits associated with a chemical provides a more suitable means of recognising growth in concern. We have two main reasons for this judgement. The first relates to how near-hits can exclude mentions of chemicals that do not relate to harm. The second relates to the limited coverage of the scientific literature achieved by Taylor and Francis publications compared to WOS.
In order to validate our theory, we plotted the two measures obtained from 1988 to 2018 for two PFAS examples that proceeded from unregulated to restricted in a relatively short time (less than 20 years). The graphs of hit counts (Figure 6.4) are reproduced in Figure 1.2 below. Each of the near-hits time series exhibits desirable features of a candidate measure of emerging concern when applied to a chemical subject to an escalating degree of concern over time: an unambiguous point after which there is a sustained increase in the measure. From this we infer that near-hits shows promise as a means of recognising when the scientific literature exhibits a sharp increase in concern regarding the harmful effects of a chemical.

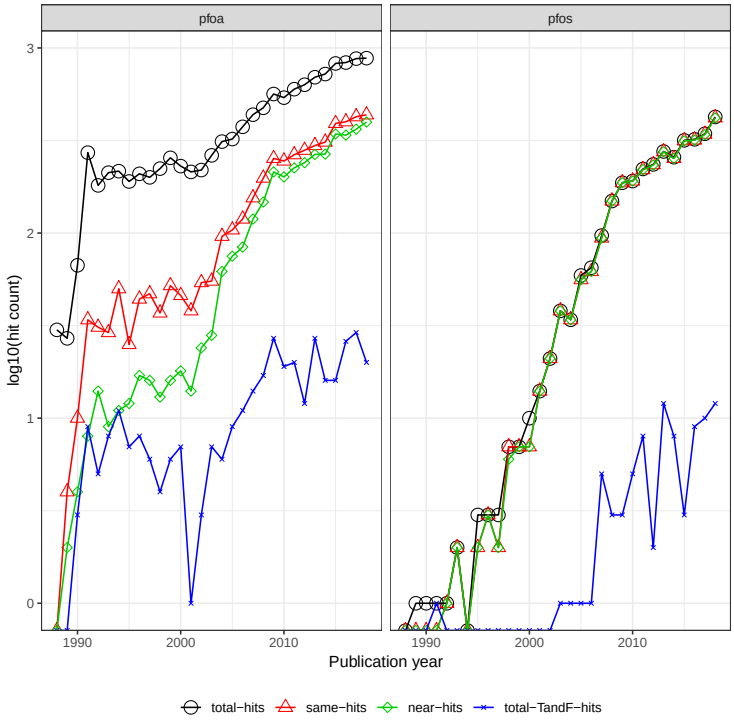

Figure 1.2.: Hit counts for PFOA and PFOS from 1988 to 2018. Note that for each chemical there is a distinct point after which the near-hits increases sharply, and continues to increase.

3. Assessment of geographical mentions:
   At the beginning of the project, CEBRA intended to obtain specific bibliographic

information associated with any hit. Yet, almost from the start we found that WOS had certain data inconsistencies that could cause database queries to incorrectly lack hits, have erroneous results, or to terminate the R code. Ongoing discussions with WOS parent company Clarivate have led to the correction of certain particular inconsistencies. Regrettably, the time taken for Clarivate to investigate and attend to our series of observations on data quality left little time to proceed with this goal.

Belatedly we determined that the keywords associated with our six sample chemicals included very few (or no) instances of major NSW population centres. As such, the sparsity of these geographical terms suggests that the current approach is not suitable. However, there are alternative approaches to the problem that we may pursue in a future project. For example, we may combine an initial search for NSW population centres with suitable terms, and attempt to filter from these mentions of specific chemicals.

4. Recommendations on next stage of project, including experimental design:
   It is appropriate to consider a larger number of chemicals, to determine whether our modelling can make similarly accurate predictions of prevalence. There is scope to experiment with more sophisticated models to see if predictions can be made more accurate. Following this, it is appropriate to compare how the methods of NSW EPA and CEBRA have classified particular chemicals, and the veracity of these classifications when compared against data.

   Similarly, consideration of the publication history of other chemicals known to have proceeded from unregulated to restricted will allow us to evaluate the reliability of near-hits as a means of detecting a spike in emerging concern.

   In this project CEBRA had limited opportunity to extract useful insights on chemicals from keywords associated with publications. We suggest that there may be value in considering this further should NSW EPA have particular associations it wants to investigate, such as those between chemicals and geographical locations, or particular health conditions.

As this was an exploratory study, time did not allow us to address certain potentially important questions, including "What is the best way to measure emerging concern?". Such questions may be considered in future studies.

2. Introduction

The New South Wales Environment Protection Authority (NSW EPA) commissioned the development of a Chemical Prioritisation Framework (CPF) in order to focus greater regulatory attention on potential risks to human and animal welfare. The framework comprises tools that assess candidate chemicals for three dimensions of risk, namely toxicity, exposure, and emerging concern. NSW EPA has engaged the Centre of Excellence for Biosecurity Risk Analysis (CEBRA) to conduct a review and validation of the components of the CPF. This report relates to the "emerging concern" dimension.

Given a particular chemical, currently NSW EPA seeks to relate the associated emerging concern to the number of publications (listed by one particular source) referring to the chemical ("hits") over successive calendar years. More specifically, publications were manually drawn from an online portal of only those works published by Taylor and Francis. For a given five-year period, NSW EPA obtained the total number of publications, and the subset of these relating specifically to "toxicology". The quotient of the latter and the former for a given five-year period yielded the associated metric for the chemical's emerging concern. Calculating this metric over rolling five-year publication periods allowed the tracking of this quantity over time. As a result of this process, NSW EPA arrived at their current metric of emerging concern: "% growth in research over the 5 year period 2013-17".

NSW EPA sought CEBRA's advice on the CPF. We proposed some extensions and refinements to the process. We investigated these new directions in a systematic manner that is reproducible and largely automated via a pilot study. In this report we will outline our approach to querying a bibliographic database in order to harvest publication data for a small number of particular chemicals.

Towards this, it is useful to discern between chemicals based on their historical prevalences. NSW EPA supplied CEBRA with a data file, listing chemicals classified into groups by their expected prevalences, as judged by expected total of distinct hits in the scientific literature from 2013 to 2017. We draw distinctions between anticipated *low* (under 100 hits), *medium* (from 100 up to 1000 hits), and *high* (from 1000 up to 10 000 hits) prevalence. We asked for these classifications to assist us in ascertaining whether large numbers of records, such as for high-prevalence chemicals, would make our intended approach to querying databases impractical, and allow us to find solutions. We randomly selected two chemicals from each of the prevalence groups (shown in Table 2.1) for further scrutiny.

The prevalence level was concealed by an uninformative text label before the transmission of data to CEBRA staff undertaking the R coding. We intended this to guard against the introduction of subconscious bias in the interpretation of results. The prevalence level was only confirmed after the analysis of results.

Given the selected chemicals, this study pursued the agreed deliverables:

1. Time series of counts of prevalences.
2. Assessment of geographical mentions.
3. Recommendations on next stage of project, including experimental design.

Table 2.1.: Chemicals (randomly) selected for this pilot study, as shown in column 2 of the file supplied by NSW EPA.

chemical name	2013–2017 Total
1,3-butadiene	231
1,3-dichloropropene	25
2,3,7,8-tetrachlorodibenzo-p-dioxin	209
fluoboric acid	27
lithium	4944
mercury	8082

With regard to Deliverable 1, one has considerable freedom in composing queries, and certain types are likely more suitable than others for this application. We will outline various query types, including one similar to that employed in the earlier CPF, which we apply to a bibliographic database. We scrutinize the counts of hits obtained so as to determine the suitability of our query types as a means of estimating emerging concern. We proceed to describe a validation experiment intending to discern whether, given a particular query type and number of chemicals, one can expect to use a time series of hits to accurately estimate near-future hits for each chemical. We intend to test whether we can expect to use prevalences in the scientific literature as a reliable input into a process such as the CPF.

Following the development of code for this part of the study, we sought to pursue a related task not previously discussed with NSW EPA that we believed could provide additional insights. We sought to determine whether we could use historical publication data to detect whenever a chemical is associated with a rapid growth in concern. We pursued this by considering publication records from our chosen bibliographic database for our sample chemicals from 1988 to 2018. We considered two measures of emerging concern, and sought to test the reliability of our approach. Towards this, we considered the publication history of two PFAS chemicals, PFOA and PFOS, known to have made a relatively rapid progression from unregulated to restricted. We proposed that if our methods were able to detect a sharp increase in concern in the historical record for each chemical, we may interpret this as a validation. We further tested the utility of our approach by comparison with the measure of emerging concern used previously by NSW EPA.

Deliverable 2 was frustrated owing to a variety of issues with the source of bibliographic data and time taken to remedy these. As a result, the time available to CEBRA to interrogate the association of our sample chemicals with geographical locations was limited. However, we found that there were few mentions of geographical locations for our sample chemicals. As such, another approach to the problem is warranted if NSW EPA wishes to pursue this task.

We delay our comment on Deliverable 3 until we have introduced the necessary background.

The remainder of this report is organised as follows. In Section 3 we present the rationale for our selection of a particular source of bibliographic data, the strengths and weaknesses of this, and potential additional information we may seek to access from the data provider should NSW EPA wish to progress beyond this pilot study. Section 4 begins with an outline of the types of queries applied to our chosen database. We

proceed to describe our validation experiment, which seeks to interrogate the use of a time series of hits in predicting future prevalence. This discussion includes our processes for modelling future hits for a given chemical, and evaluating whether we have made accurate predictions. We present results for our six randomly-selected chemicals in Section 5, comment on the accuracy of our prediction process, and compare our classifications of chemicals against those obtained by using Taylor and Francis hits.

In Section 6 we present our approach to detecting rapid growth in the concern relating to the harmful effects associated with a chemical from a bibliographic database, and the validation of this approach. Section 7 presents the unsuccessful attempt to associate our sample chemicals with geographical locations in NSW, and proposes an alternative approach to the problem. With Deliverable 3 in mind, we draw conclusions and make recommendations in Section 8. In Appendix A we demonstrate the reliability of our database scraping code by showing that hits obtained agree with those obtained by querying our bibliographic database via typing queries into a web browser.

3. Data source considerations

We sought a source of publication data that we expected would:

- allow comparison of our results against those from NSW EPA's CPF, (which was limited to Taylor and Francis publications),
- have a broad coverage of the scientific literature that we expected to be relevant,
- provide a flexible search syntax that would allow us to experiment with various types of queries,
- allow us to scrape a significant amount of bibliographic information in a manner that could be largely automated,
- allow us to readily conduct validation of the results obtained by our code against results obtained directly from some other method (e.g. a web browser) in a manner that did not require time-consuming manual processing.

Given our access to Clarivate products and product discussions with the company, we decided to investigate the Web of Science (WOS) API. We note that while other APIs may have also been appropriate for our purposes, progress with these was impeded by limited access or slow responses to requests for further information or access.

In addition to those points listed above, some particularly useful features of the WOS API are:

- A "Topic search', which allows the user to simultaneously search the fields "title", "abstract", "keywords" (supplied by author), and "keywords plus" (which Clarivate advises has content "...supplied by an algorithm that provides expanded terms stemming from the record's cited references or bibliography").

- The search syntax has features not present in other APIs. For example, the WOS API allows queries such as "term1 SAME term2" to find occurrences of term1 and term2 in the fields relevant to a Topic Search, and "term1 NEAR/n term2" can similarly find occurrences of term1 within (some positive integer value) n words of term2.

Notwithstanding the various agreeable features of the WOS API, scrutiny over the course of this study has made us aware of certain limitations. We found these particularly in early attempts to obtain specific information from each hit. We found that scraping code could exit prematurely upon encountering a page that should have some number of publications, yet actually had none. Similarly, publication data stored in an incorrect format could cause CEBRA's otherwise-reliable code to terminate on retrieval.

CEBRA and Clarivate have had an ongoing dialogue on such data issues. Over the course of this project, Clarivate have remedied a number of specific concerns. At the time of writing this report, the current outstanding issues are relatively minor. For example, we have seen that a query which specifies a particular range of publication years can return some small number of hits that do not match the search criteria. It

appears that these erroneous hits are merely inconvenient, as they are readily filtered out in pre-processing of results prior to our analysis. We have reported a number of cases of erroneous hits to Clarivate, who are investigating the matter.

4. Data extraction methodology, initial results, and a process for validating modelling of future interest

Any individual query used in this study relates only to one particular chemical. In Section 4.1 we present queries subject to differing degrees of selectivity. We outline the application of these to our data source in Section 4.2. The results returned by this application to the chemicals shown in Table 2.1 are shown in Section 4.3. A comparison of results allowed us to gain an appreciation for how the query type influenced the order of magnitude of hit counts obtained for our sample chemicals. We used this in selecting a query type expected to provide reliable results in the subsequent validation exercise, outlined in Section 4.4. We describe some of the choices made in this process in Section 4.5.

In forming queries, care was taken to ensure that the query was an accurate representation of our intent. For example, when a chemical name consisted of multiple words, the name was enclosed in quotation marks so as to return only that exact phrase, rather than hits which merely featured each of the words. Other precautions related to ensuring that queries were implemented in "url encoding" so as to be interpreted properly by the WOS API.

4.1. Query types: an introduction to four types of database "hits"

Each query applied to WOS was composed of multiple conditions. We define a "hit" as a publication which satisfies all conditions of a query.

The simplest query for a chemical of interest was for "**total-hits**". The component conditions of the query (combined with AND) are:

T1 a "Topic Search" for all supplied alternative names (combined with OR; a publication is retained as a possible hit if it contains at least one of these alternatives),

T2 a search of the "Publication Year" field for some value(s) from a range, initially, 2012-2018 inclusive.[1]

We also required a means of comparing our results against a previous study commissioned by NSW EPA, which considered only those works published by Taylor and Francis. Subsequently we refer to these as "**total-T&F-hits**". A query to return total-T&F-hits (with components combined by AND) has components:

T&F1 the "total-hits" conditions (T1 and T2 above),

[1] We will explain this further in Section 4.2.

T&F2 the requirement that a hit has a digital object identifier (DOI) beginning with the Taylor and Francis prefix of *10.1080*.

We note that total-hits or total-T&F-hits may count unsuitable publications — those unrelated to an adverse effect on humans or the environment. As such, total-hits or total-T&F-hits may not be the most appropriate measure of prevalence for the CPF for all situations. In order to address this, we implemented more restrictive queries by requiring that any hit must match features T1 and T2 above, as well as further conditions ("modifiers") in the WOS "Topic Search" field. We intended these modifiers to relate to adverse effects.

A search for "**same-hits**" added to conditions T1 and T2 the further condition that at least one modifier must appear in the same field as a form of the chemical name. The "**near-hits**" further required that at least one modifier occurred at most three words from where a form of a chemical name appears in a field.

The modifiers used in the same-hits and near-hits queries are presented in Table 4.1. We note that where "*" is shown, this denotes a "wild card" operation. To illustrate this concept, a query with "*toxic*" in the Topic field will search for all words containing "**toxic**", such as **toxic**ity or eco**toxic**ology.

Table 4.1.: Modifiers in the queries intending to return "same-hits" and "near-hits".

Modifier	carcinogen∗	dangerous∗	fatal*	harmful*
	injurious	lethal∗	noxious*	poison∗
	∗toxic∗	unsafe		

4.2. Query applications

Almost immediately, the unexpected data issues outlined in Section 3 prevented us from obtaining geographic information from hits returned, as we had planned. The time taken for rounds of bug reporting and their correction allowed little opportunity for us to return to this task within the time allowed for this project. For now, we make only brief remarks in Section 7. However, we expect that the R code we developed to obtain details of individual publications may be profitably applied in a future project.

As an alternative means of obtaining publication counts, we implemented a "shallow scrape". This applies a query to the database for each (calendar) year in the nominated publication range. This allowed us to obtain counts of publications for a given year directly, and does not attempt to scrape any further bibliographic information.

The shallow scrape has the benefit that it is relatively quick to run — scraping results from 2012-2018 (inclusive) for our sample of six chemicals were obtained in under six minutes on a 2017 laptop (2.3 GHz Intel Core i5, two processors). This allowed us to trial different types of queries in an efficient manner. We note that we have not performed substantial code optimisation up to this point. This will be a priority for any future work seeking to consider a larger number of chemicals.

4.3. Initial results for the four types of hits

In Figure 4.1 we show the total-hits, same-hits, near-hits, and total-T&F-hits (each on a base-10 logarithmic scale) obtained for the six chemicals under consideration, and for each year from 2012 to 2018.

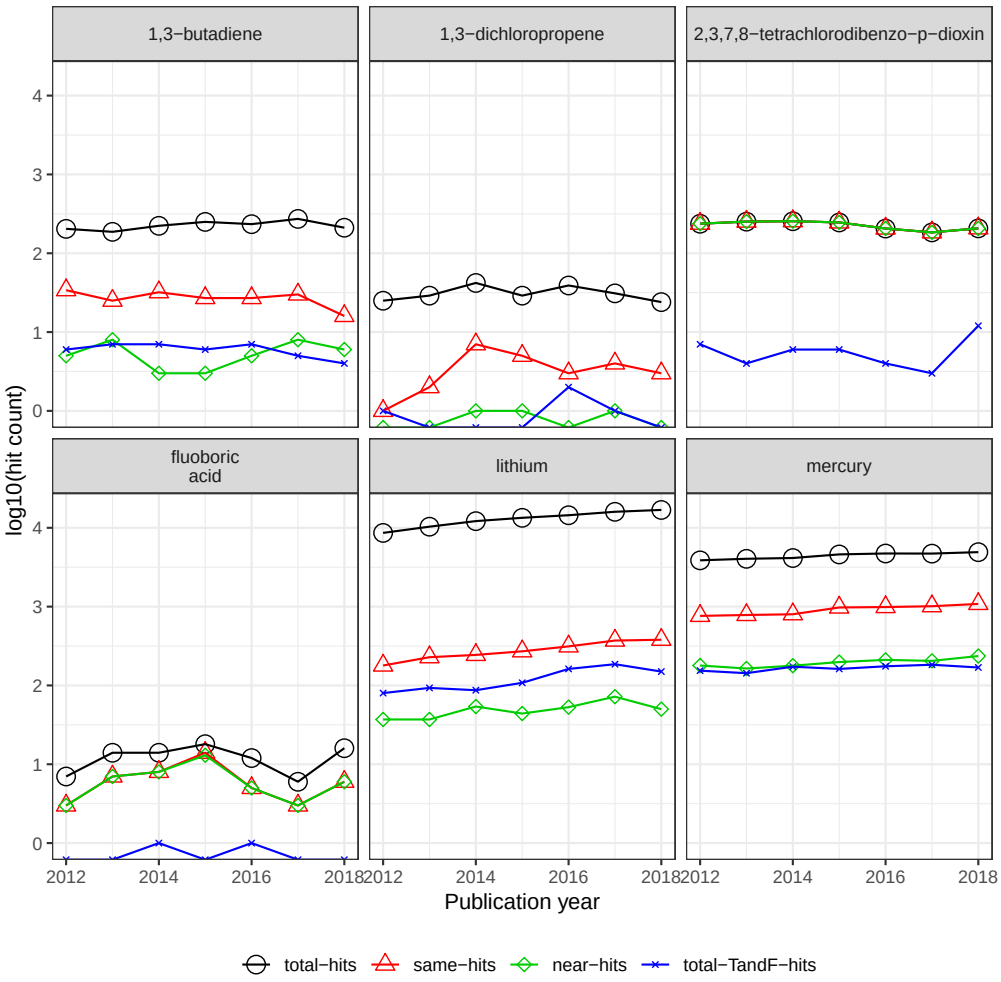

Figure 4.1.: Time series of the four hit types in WOS for the sample of chemicals from 2012-2018. Whenever fewer than four series are shown on a panel, some series are overlaid exactly by the "near-hits" series.

We note that:

- As expected, for all chemicals and years, total-T&F-hits are less than total-hits.
- for all chemicals except 2,3,7,8-tetrachlorodibenzo-p-dioxin, the total-hits in a year exceed (possibly greatly) the corresponding same-hits or near-hits values.
- There is no consistent ordering of the total-T&F-hits relative to the near-hits across calendar years and chemicals. We suggest that this makes total-T&F-hits an unreliable means of judging prevalence of a chemical in the scientific literature, potentially making this measure unsuitable for the CPF.
- The total-hits for a chemical may be quite similar to the associated near-hits (as for 2,3,7,8-tetrachlorodibenzo-p-dioxin), or substantially greater (as for lithium). This suggests that total-hits overestimate emerging concern. Thus, it is appropriate to proceed with a more restrictive query than that used to obtain total-hits.

4.4. Outline of the validation experiment

As this study seeks to associate chemical substances with harmful effects on humans and the environment, the lack of specificity of total-hits and total-T&F-hits makes these unsuitable. Further, it is possible that some same-hits returned for a given chemical are spurious, such as when some modifier in an abstract appears some lines away from the chemical of interest, and does not relate to it. We expect that the stricter requirement imposed by the near-hits search should eliminate a substantial proportion of spurious hits. Whilst allowing for up to three words between a chemical and a modifier may produce some false hits, this flexibility does allow for us to detect alternative means of reporting on a chemical (e.g. "...recorded harmful chemical X concentrations ...", " ...after 12 months, X persists at toxic levels ..." so that true hits are not excluded. As such, given the importance of excluding false hits yet retaining true hits as much as possible, we deemed the counts of near-hits to be the most appropriate data for our purpose of predicting future mentions of harm caused by a chemical in the scientific literature.

For each individual chemical in Table 2.1, we proposed the following approach to estimating a future value of emerging concern:

1. Perform a shallow scrape to obtain counts of hits from 2012-2018, inclusive.
2. From this, select 2012-2016 as calibration data and 2017-2018 as validation data.
3. Fit a "trend model" to the calibration data.
4. Use our fitted model to produce some predictor of the values in our validation data.
5. Determine whether or not predictions from our process could lead to "acceptably close" predictions of the recorded 2018 "near-hits".

Suppose in using this process we obtained a positive answer to 5. for each chemical. We may regard this as indicating that the process of steps 1. to 4. provides a reasonable way to anticipate the interest of the scientific literature in a chemical two years into the future. That is, we can have some security that our approach to predicting near-hits will be reasonable in the case where we do not have validation data.

Suppose we use our predictions of 2018 hits to assign chemicals into groups of high, moderate, and low prevalence. We propose that those chemicals judged as high prevalence are potential priorities for NSW EPA's monitoring efforts.

We may also apply the process outlined above in a different manner. Suppose in using the process we obtained a negative answer to 5. for a chemical. When an observation is sufficiently large so as to exceed our prediction and is beyond the range of "acceptably close", we may interpret this as a sign of an "uptick" in the interest in that chemical. We may use this to decide that the chemical is deserving of further attention.

4.5. Choices made in modelling

The process outlined above is subject to certain choices. These may be reviewed should CEBRA consider a greater number of chemicals subsequently.

We chose a five-year duration for the calibration data set as we expected this to mitigate potential drawbacks of other choices. A longer calibration period would increase the programme run time. Further, this could conceivably give too much weight to

counts prior to when a chemical became associated with health risks, leading to underestimates of future hits. Conversely, a shorter period may place too much weight on recent history and be unable to recognise an emerging trend in hits that has flattened only in the recent past.

As our "trend model" for any time course of hits we used an ARIMA model where R chose the model order. For each chemical, we use the calibrated model to predict the near-hit count for the latest year of interest (2018). We use this estimate in labelling a chemical as either low, moderate or high prevalence.

In comparing near-hits predictions and data, we use "acceptably close" to mean that the 95% prediction interval for the 2018 near-hit count contains the 2018 datum. We note that other significance levels may be used for prediction intervals, where the choice of these will reflect differing appetites for risk.

We present the results of the validation experiment in the next section.

5. Validation experiment results and caveats

The ability of some trend model to approximate the actual 2018 near-hit value for a given chemical (recall predictions are only made for 2017-2018) is shown in Figure 5.1. Additionally, features of the model fit to some hits time series are shown in Table 5.1. In four of six cases, the 2018 datum is within the 95% prediction interval. We can disregard the 1,3-dichloropropene case as the calibration data has years with zero hits, and hence we cannot predict the 2018 hit count. We do not expect this chemical to be a monitoring priority, and would set the 2018 hit count to zero. In the final case (mercury), the actual 2018 near-hit value is slightly larger than the upper bound of the 95% prediction interval. This inability to capture the uptick in the validation series may demonstrate a limitation of our simple modelling approach.

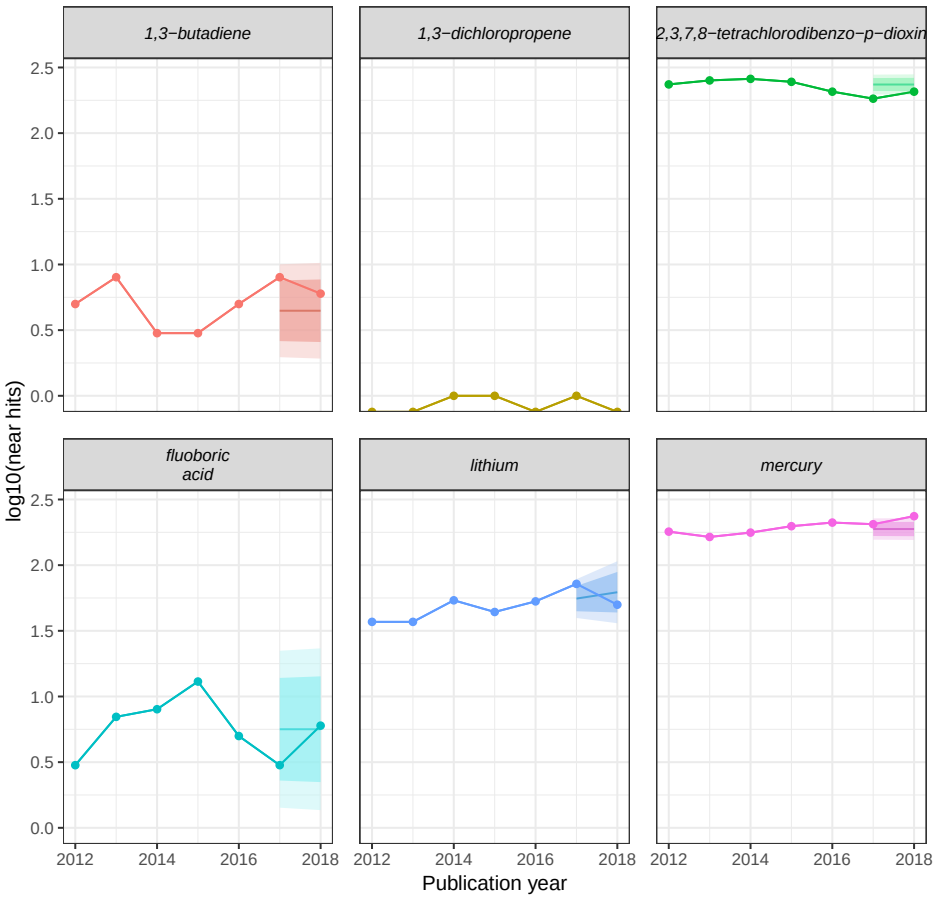

Figure 5.1.: Plot of near-hits data from 2012-2018 for the six selected chemicals (lines with points), and for the 2017-2018 data, point predictions (solid lines) and prediction intervals (lighter shading for 95%, darker shading for 80%).

Table 5.1.: A summary of model fitting to the calibration set and the model's performance on the validation set for each of the selected chemicals. A distinct model is fit to the time series of counts for each chemical. Displayed for each chemical are quantities derived from the model fitting process: SSE (sum of squared errors) for the model fit to the calibration dataset, σ^2 is the variance of the residuals from the model fit, Lower 95% PI and Upper 95% PI are the lower and upper 95% prediction intervals respectively, and Estimate is the model's estimate of the 2018 count value (2018 datum). We use Fit quality to summarise how the 2018 datum compares against the 95% prediction interval. If the datum is contained by the prediction interval, we label our process for predicting future counts based on the calibration data set as "ok" .

	Calibration summary		Validation summary				
Chemical	SSE	σ^2	Lower 95% PI	2018 datum	Estimate	Upper 95% PI	Fit quality
1,3-butadiene	0.163	0.033	0.283	0.778	0.648	1.012	ok
1,3-dichloropropene	NaN	NaN	NaN	-Inf	NaN	NaN	NaN
2,3,7,8-tetrachlorodibenzo-p-dioxin	0.007	0.001	2.293	2.316	2.370	2.448	ok
fluoboric acid	0.465	0.093	0.135	0.778	0.751	1.366	ok
lithium	0.028	0.006	1.558	1.699	1.794	2.030	ok
mercury	0.009	0.002	2.191	2.373	2.275	2.358	underestimate

Figure 5.2 presents a comparison of the base-10 logarithm of actual 2018 "near-hit" values for our six chemicals against our model-derived estimate of these values. For each case, the 95% prediction interval is shown with vertical bars. (No result is shown for 1,3-dichloropropene owing to the lack of hits noted above.) The solid black (diagonal) line indicates the theoretical case in which data agrees exactly with predictions. Our prediction process is able to resolve the sampled chemicals into distinct groups; 1,3-butadiene, 1,3-dichloropropene, and fluoboric acid are of lowest prevalence, lithium is of intermediate prevalence, and 2,3,7,8-tetrachlorodibenzo-p-dioxin and mercury are of highest prevalence.

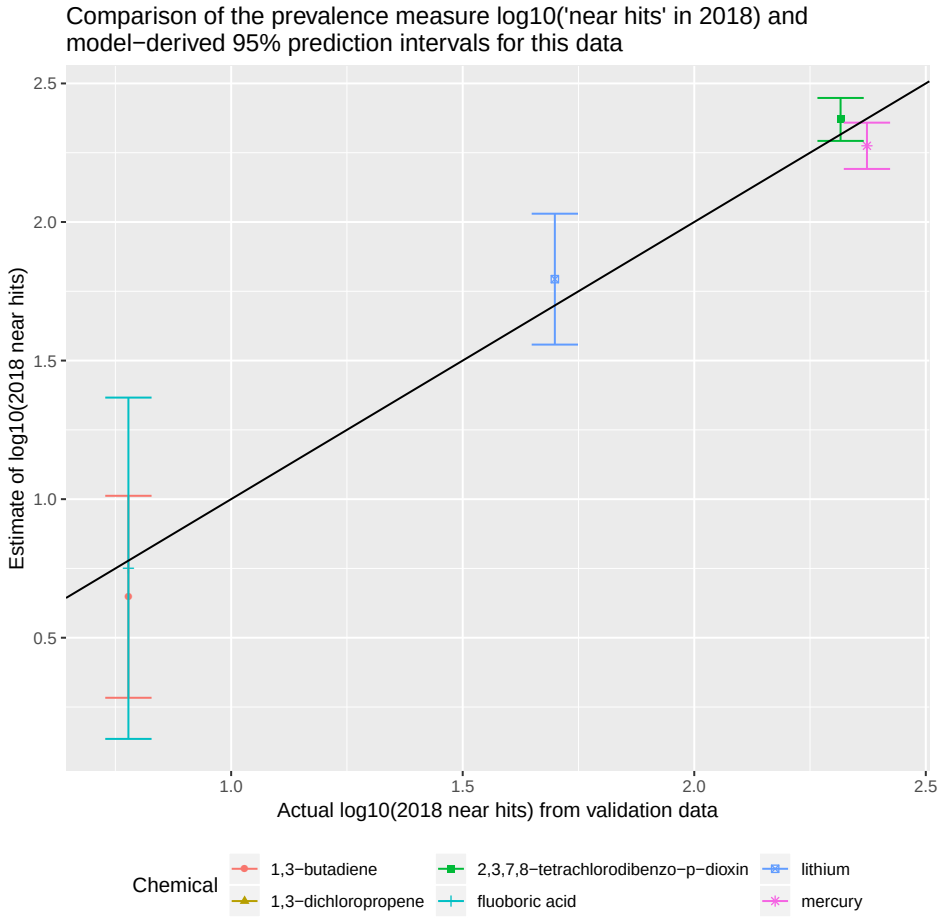

Figure 5.2.: A comparison of predictions of the 2018 "near-hit" value and the actual 2018 datum for each of the selected chemicals. The black line shows the theoretical case where predicted and actual values are equal.

We will now formally assign the chemicals considered to prevalence categories, by adapting the categories defined in the Introduction. As those categories were defined in terms of totals over a five-year period, here we will divide the thresholds by five so that we may classify a chemical via our estimate of its 2018 near-hits count. In the interests of comparison, we use the total-T&F-hits in the same way to classify our chemicals. These results are shown in Table 5.2. We note that we have set our classification approach using near-hits the task of predicting prevalence two years into the future, which is a more stringent test than that faced by the use of total-T&F-hits.

Table 5.2.: A comparison of measures of prevalence of our six chemicals derived from the time series of near-hits and total-T&F-hits, and allocations of chemicals to a prevalence category based on these. Most notably, the assessment of prevalence for 2,3,7,8-tetrachlorodibenzo-p-dioxin using near-hits gives a far greater estimate than that obtained using total-T&F-hits. For T&F-hits, allocation is made by reference to an average of hits over the five-year period 2013-20217. For near-hits, allocation uses a transformation of the trend model's prediction of the base-10 logarithm of the 2018 near-hits value. Estimated values are given to two decimal places. The use of ✗ indicates that no prediction is possible owing to years of zero hits in the count history for a chemical — we interpret this as 0.

	Near-hits		Total-T&F-hits	
Chemical	2018 data/category	est. 2018 data/category	2018 data/category	av. 2013-2017/category
1,3-butadiene	6 / low	4.44 / low	4 / low	7.6 / low
1,3-dichloropropene	0 / low	✗/ low	0 / low	0.80 / low
2,3,7,8-tetrachlorodibenzo-p-dioxin	**207 / high**	**234.50 / high**	**12/ low**	**6.00 / low**
fluoboric acid	6/ low	5.63 / low	0/ low	0.40 / low
lithium	50 / moderate	62.21 / moderate	150 /moderate	174.60 / moderate
mercury	236/high	188.36 / moderate	169/ moderate	231.00 / high

Considering the near-hits results, in the first five cases we see that the classification arrived at by modelling the 2018 datum is the same as that we would obtain by using the datum directly. However, for the case of mercury we see that a (simple) time series model classifies the chemical as being of moderate prevalence, whereas the data indicates that it is of high prevalence. This may indicate an uptick in the interest relating to mercury. Alternatively, the underestimation may signal a need to develop our time series models, such as by giving more weight to more recent values in the fitting to calibration data. However, at this exploratory stage, the result may suggest the value of considering more cases so as to ascertain whether our near-hits metric is a reliable indicator of prevalence.

Similarly, further test cases will allow us to evaluate the suitability of a five-year calibration data set. This modelling decision appears reasonable for the chemicals considered in Figure 5.1, however, none of these exhibited a rapid growth in concern in the scientific literature. As such, it is appropriate to validate our methods further by modelling the hit counts obtained for such chemicals in a future study.

The results for the total-T&F-hits are somewhat similar to those obtained for the near-hits results. We see the classifications applied to the first five chemicals based on an average of near-hits agree with those classifications resulting from the 2018 total-T&F-hits.

We see a striking difference between the near-hits results and total-T&F-hits results for 2,3,7,8-tetrachlorodibenzo-p-dioxin. The near-hits (modelled and data) are substantially larger than the total-T&F-hits, leading to the chemical being classified as high prevalence under the former measure, and low prevalence under the latter. This result suggests that total-T&F-hits is not guaranteed to capture a substantial volume of interest in a chemical. This justifies the further consideration of measures of prevalence.

Other differences in classifications obtained using the two measures relate to mercury. Table 5.2 shows that the near-hits modelling classifies mercury as of moderate prevalence, yet the actual 2018 near-hit count suggests a classification as high prevalence. That is, the near-hits approach leads to an underestimate of mercury prevalence. In the total-T&F-hits case we see an overestimate of mercury prevalence, as the metric suggests a "high" categorisation whereas data indicates "moderate". We may have to scrutinize details of the actual hits in order to ascertain whether an approach based on near-hits is a reliable indicator of prevalence. However, the substantial underestimate of prevalence for 2,3,7,8-tetrachlorodibenzo-p-dioxin using total-T&F-hits compared to that obtained using near-hits suggests that there is a risk of misclassification of chemicals using an approach based on Taylor and Francis publications.

We note that for the purposes of this pilot study, we have taken only a simplistic approach to the modelling of hit counts, and yet found the process adequate for five of our six randomly-selected cases. As a result, we expect that our approach to predicting future counts as part of a risk prioritisation is essentially sound. We acknowledge that it may be possible to improve the process as a wider range of chemical substances are considered. We expect our results to improve should we require the prediction of only the next year's prevalence for chemicals, rather than for two years ahead.

Figure 5.1 shows that near-hit counts do not change greatly over a period as short as two years. Further inspection of the figure (or Figure 4.1) shows that in each case the 2016 near-hit value is approximately equal to the 2018 value. As such, we may expect that a chemical classified as belonging to one prevalence group has (qualitatively) only a small probability of changing groups. We will need to scrutinise the behaviour of

other chemicals in order to rigorously test this hypothesis.

Based on the results collated here, we would expect that :

- 1,3-butadiene, 1,3-dichloropropene, and fluoboric acid, previously considered low prevalence, do not change status.
- lithium, previously considered of moderate prevalence, does not change status.
- 2,3,7,8-tetrachlorodibenzo-p-dioxin, and mercury, previously considered of high prevalence, do not change status.

Inspection of the data shows that each of these predictions are accurate. This warrants further inspection on a broader range of chemicals. It also supports our note earlier in this section that it may be appropriate to give more weight to more recent points in time series modelling of future hits.

In the next section we discuss the use of WOS to determine when the scientific literature begins to take a greater interest in the harmful effects of some chemical, which we interpret as a measure of emerging concern for that chemical.

6. Detection of chemicals of emerging concern from publication records

In the previous section we considered predicting future interest in a chemical based on its historical publication record. NSW EPA aspires to a related objective: a process for using such publication records to detect those chemicals which show a sharp increase in some measure of related research interest (trend recognition). For such chemicals, the increase is a signifier of "emerging concern". As such, the relevant chemicals may become candidates for increased attention.

We recall that previously NSW EPA used as its measure of emerging concern for a given chemical a quotient of counts drawn from Taylor & Francis publications: the average number of annual hits associated with toxicology over a five-year period divided by the average of the total annual number of hits drawn from the same period. We will examine the usefulness of such a strategy by employing a similar measure that we may derive from WOS data — the quotient of near-hits and total-hits in a given year. We choose to not use average values owing to the undesirable possibility that averaging may cause a lag in the time taken for the quotient to show a notably high (and thus potentially instructive) value.

We have adapted the code developed for the prediction exercise to provide input to the complementary task of trend recognition. Following advice from NSW EPA, we consider the extended publication range of 1988 to 2018. We also consider the types of hits introduced earlier over this publication range. We present results in this section, and a validation of our approach.

6.1. Consideration of measures of concern for the six sample chemicals

We begin our investigation via a consideration of the different types of hits for the six sample chemicals. These time series are shown in Figure 6.1. Themes observed for the 2012-2018 time series shown in Figure 5.1 are also in evidence here; the tendency of total-T&F-hits to take substantially smaller values than other types of hits, and, the tendency for time series to show a modest upward trend (or no particular trend) rather than sharp (instructive) increases.

A consideration of the quotient of near-hits and total-hits for our sample chemicals is shown in Figure 6.2. The graphs show that in five of six cases, the quotient of hits for a given chemical tends to maintain a low or high value. The quotient for fluoboric acid shows some rapid increases as part of its oscillatory behaviour. However, when judged against the small number of hits for this chemical, this provides only limited evidence of emerging concern.

A weakness of this demonstration is that our randomly selected chemicals are not known to show a spike in emerging concern that we could expect to capture with

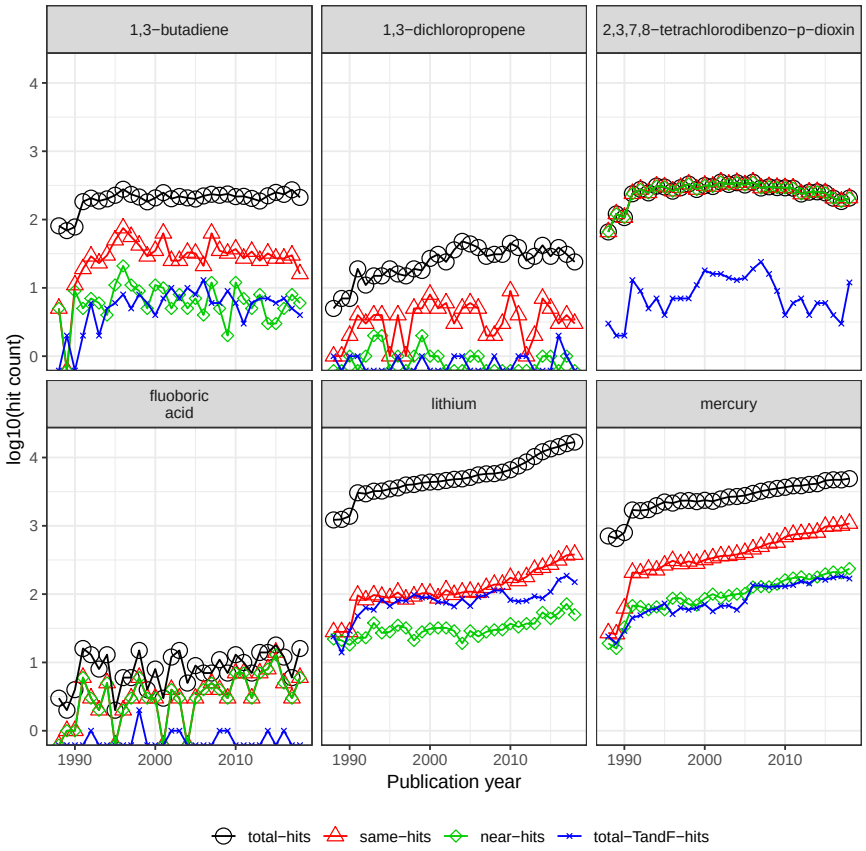

Figure 6.1.: Hits from 1988 to 2018 for the six sample chemicals.

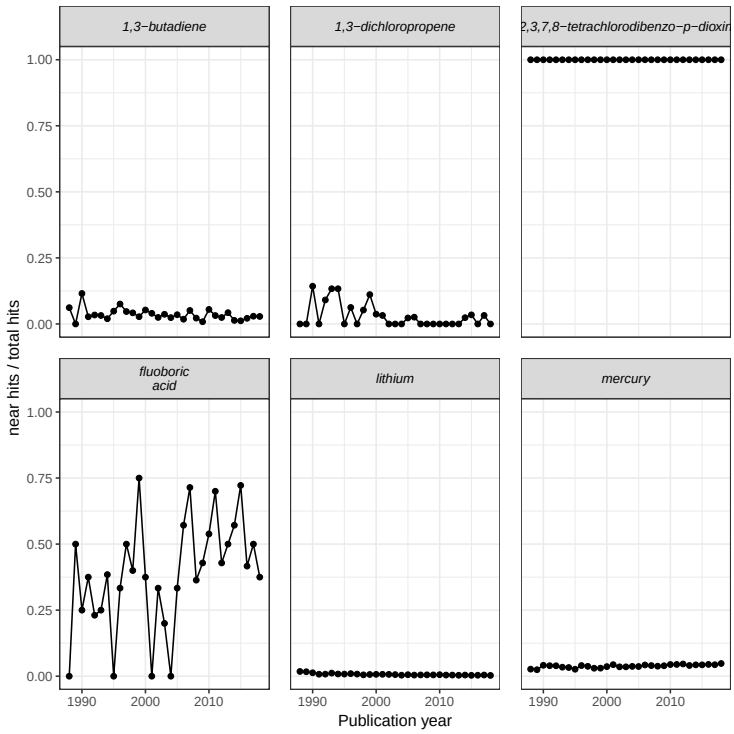

Figure 6.2.: The quotient of near-hits to total-hits for 1988–2018 and each sample chemical.

our methodology. To remedy this, we will consider the ability of our method to recognise two demonstrated cases of chemicals for which regulators progressively increased their concern, leading to restrictions within a relatively short time after chemicals were judged as harmful. Should our methodology recognise a point at which concern grows rapidly, we may interpret this as a validation of the essentials of the process employed here. Alternatively, a failure to recognise emerging concern will indicate the need for further experimentation.

6.2. Validation of metrics against two "classic" PFAS examples

We seek to validate our methodology by its application to two PFAS chemicals — PFOA and PFOS. Regulators developed a substantial increase in concern regarding the effects of these chemicals in a relatively short time. We can judge this, for example, from the content of alerts from Australia's National Industrial Chemicals Notification and Assessment Scheme (NICNAS) over time. NICNAS first issued alerts relating to PFOS in July 2002, and to PFOA in November 2004. These alerts summarised the ongoing or draft assessments of the PFAS in question undertaken by international regulators. By February 2007 regulatory concern towards each chemical had become much more definite. One NICNAS alert advised industry to " ...seek alternatives and seek to phase out PFOA ...".[1] Another alert advised that PFOS be restricted "...to essential uses where no suitable less hazardous alternatives are available."[2] This escalation of concern for chemicals already in widespread use led NSW EPA to cite PFOA and PFOS as two examples of the motivation for this study. We will apply our methods to these chemicals in order to ascertain whether or not we may discern a rapid increase in associated concern.

To aid evaluation of our results, in Figure 6.3 we reproduce emerging concern results for chemicals including PFOA and PFAS determined using NSW EPA's previous metric. Given the general tendency (noted above) of T&F hit counts to be substantially smaller than counts for other hits, it is only appropriate to make qualitative comparisons between our results and those in Figure 6.3. In a similar manner as for the previous section's chemicals, for PFOA and PFOS over the publication range from 1988 to 2018, we consider the time courses of hits (Figure 6.4), and the quotient of near-hits and total-hits (Figure 6.5).

Figure 6.4 shows that for PFOA we can discern a sharp and sustained increase in the near-hits value from 2001 to 2009, with the increase continuing (albeit at a reduced rate) to 2018. We also note that the T&F hits are substantially lower, which underestimates the research interest. The graph for PFOS shows a similar sustained increase in near-hits, from 1997 to the end of the series. We note that the near-hits behaviour differs from the rise and fall of the concern measure in Figure 6.3, making the former a more consistent signifier of concern regarding PFOA and PFOS from (say) 2001 onwards.

[1] https://webarchive.nla.gov.au/wayback/20091030155531/http://www.nicnas.gov.au/Publications/NICNAS_Alerts/EC_Alert6.pdf, last accessed December 2 2019.

[2] https://webarchive.nla.gov.au/wayback/20091030155531/http://www.nicnas.gov.au/Publications/NICNAS_Alerts/EC_Alert5.pdf, last accessed December 2 2019.

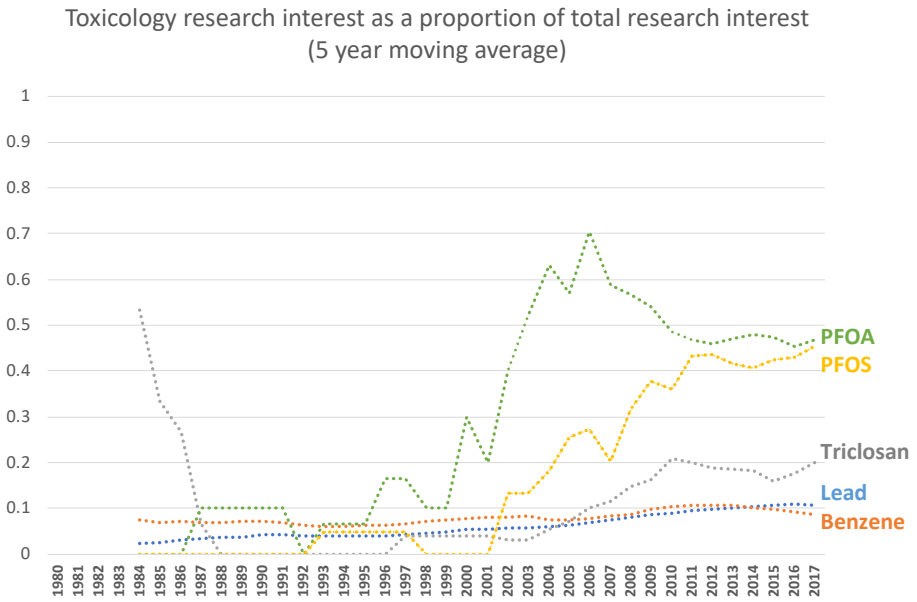

Figure 6.3.: A graph (commissioned by NSW EPA) showing the original measure of emerging concern for a selection of chemicals.

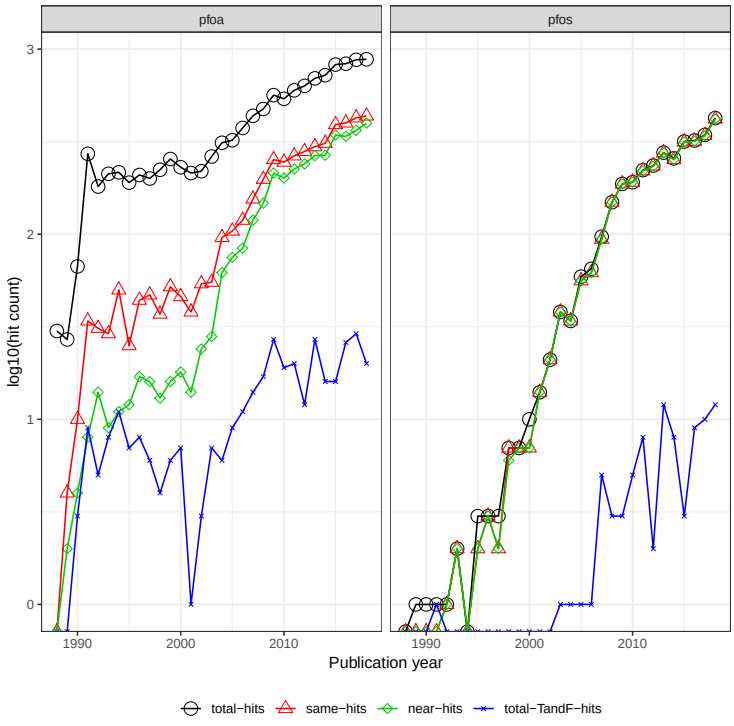

Figure 6.4.: Hit counts for PFOA and PFOS from 1988 to 2018.

The quotients of near-hits and total-hits in Figure 6.5 do not make the same clear statement. The PFOA graph shows an upward trend from 2001, but also shows a plateau (2009-2014) that is not seen in the corresponding graph in Figure 6.4. This feature obscures the actual growth in research interest relating to potential harm caused

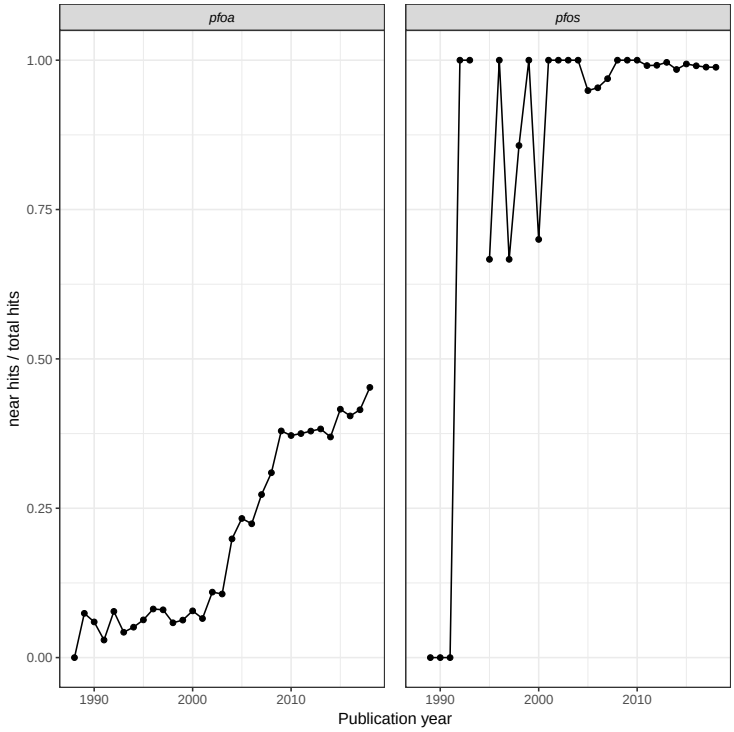

Figure 6.5.: The quotient of near-hits and total-hits for PFOA and PFOS for the publication year range from 1988 to 2018.

by the chemical. The PFOS graph shows an abrupt jump from zero to one, some oscillations around a high value (approximately 0.85), and then it stabilises around one for the rest of the series.

These observations suggest weaknesses of the quotient metric. In the PFOA example, increases in the denominator can obscure the effect of increases in the numerator. The PFOS case provides an example of how small values for total-hits and comparably small values for near-hits can cause a large initial value that does not change greatly over time. This feature may cause an overestimate of concern for some chemicals.

We propose that a quotient of quantities is not an ideal metric for emerging concern, as near-hits alone may be more informative. For example, we may deem a chemical to be worthy of greater scrutiny once its near-hits satisfies some "threshold condition", such as exceeding 30 (alternatively, 100) in a year for two consecutive years; approximately 1.5 (alternatively, exactly 2) on our log10 scale. In the PFOA case, Figure 6.4 shows that the threshold condition is satisfied in 2005 (alternatively, 2008). Similarly, for PFOS, the condition is satisfied in 2004 (alternatively, 2009).

From these examples we propose that a measure such as near-hits may provide an adequate indicator of when the scientific literature has recognised some chemical as a matter for concern. However, we would need to test this proposition by considering of a greater range of chemicals, including some that have shown the PFAS-like evolution from unregulated to restricted, and others which continue without regulation, to ascertain if we can reliably discern between such groups of chemicals.

We conclude this section with a final note on the choice of metric used to evaluate emerging concern. We note an aspect of the discussion in: "3M Knew About the Dangers of PFOA and PFOS Decades Ago, Internal Documents Show" by Sharon Lerner,

published in The Intercept, August 1 2018:[3]

> Since 2000, the number of scientific articles published on the health effects of PFAS has increased more than tenfold. The findings have linked the chemicals to a wide range of health effects in people, including testicular and kidney cancer, obesity, impaired fertility, thyroid disease, and the onset of puberty.

From this extract we suggest that further pursuing the association of chemicals and specific (human) "health effects" in scientific articles may provide a useful contributor to the process of judging "emerging concern".

[3] https://theintercept.com/2018/07/31/3m-pfas-minnesota-pfoa-pfos/, last accessed November 3rd 2019.

7. An investigation of publication keywords, including geographical locations

For each of our six sample chemicals, we conducted a search of WOS for publications associated with selected geographical terms, and sought to interrogate the keywords returned for these hits. Searches for publications mentioning the population centres: Albury, Bathurst, Central Coast, Coffs Harbour, Dubbo, Lismore, Maitland, Newcastle, Nowra, Port Macquarie, Richmond, Tamworth, Wagga Wagga, and Wollongong, and New South Wales, did not yield any results. A search of keywords for Australia returned two hits for lithium and four for mercury. A search for Sydney returned only one hit for mercury.

These results suggest that first searching publication data for mentions of chemicals, and then interrogating the keywords of hits returned, may not be a viable means of obtaining associations between chemicals and geographical locations.

Regardless, the keywords and phrases associated with publications mentioning the six sample chemicals contain a variety of terms. It is possible that further scrutiny of these in a future study will provide associations that will direct future research. We provide an example of the most frequently occurring keywords and phrases for one of our sample chemicals in Figure 7.1.

Given the sparseness of geographical mentions of interest to NSW EPA, it may be prudent to modify the search process such that we begin by considering publications associated with a geographical location. For example, suppose we require any WOS Topic Search query to include "Sydney OR New South Wales" and variants on "pollutant" or "contamination". We may then proceed to determine the frequency of mentions of chemicals in the hits returned over publication years, which may show emerging concern for certain chemicals. In order to demonstrate the type of results obtainable from the query proposed, we present results from WOS for 1998 to 2018 in Figure 7.2. As for other searches described in this report, we expect that the validity of the search terms employed to play a large role in the usefulness of search results. We welcome input from NSW EPA on such matters.

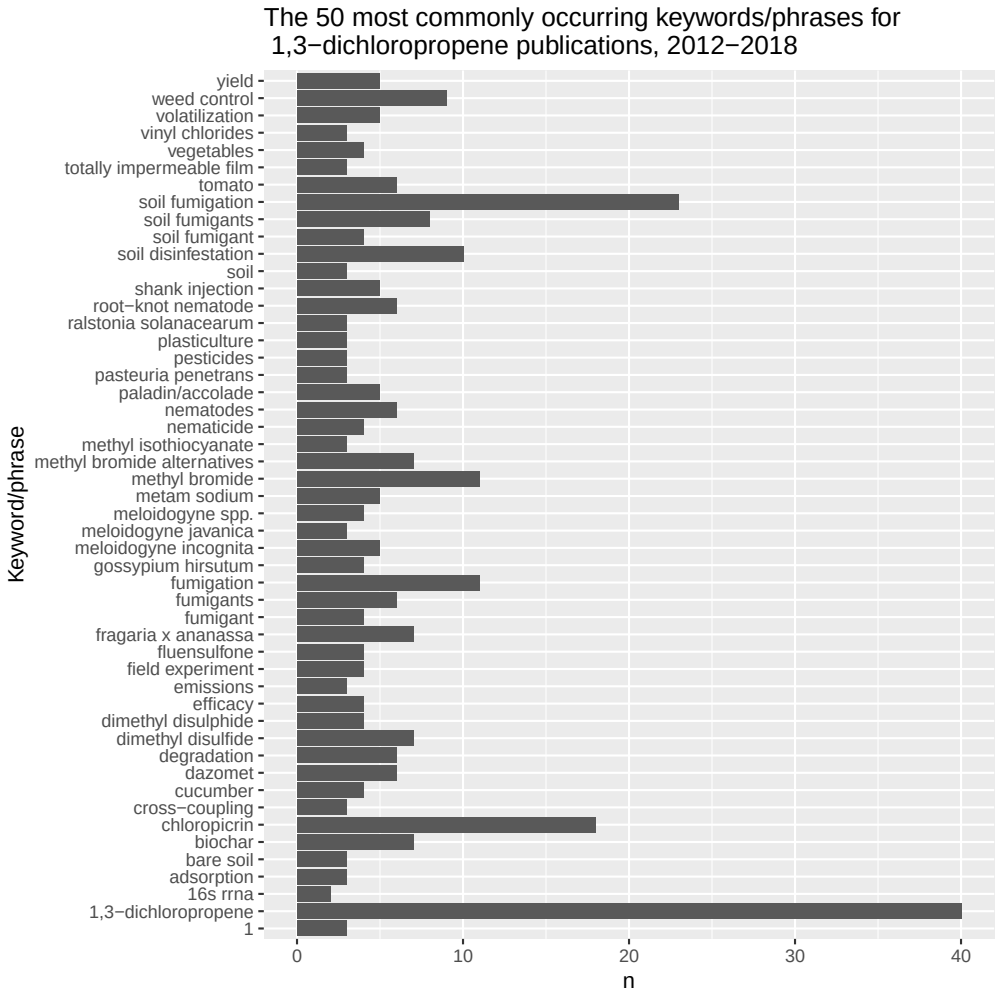

Figure 7.1.: The 50 most commonly occurring keywords and phrases from publications including 1,3-dichloropropene, to illustrate the breadth of keywords which may bear further scrutiny.

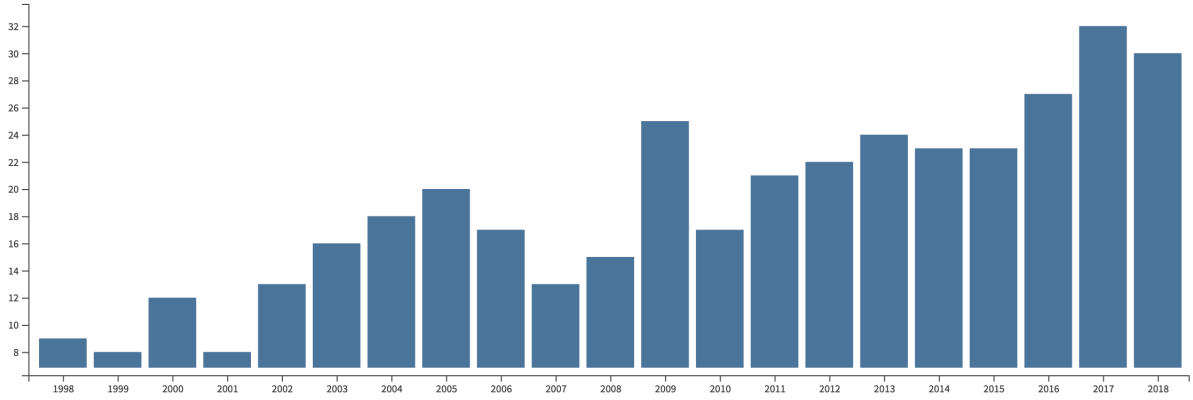

Figure 7.2.: Time series of hits returned for a WOS Topic Search for (Sydney and "New South Wales") AND (contaminant* OR pollutant*) from 1998 to 2018, to demonstrate one means of obtaining location-specific publication records for further scrutiny.

8. Promising avenues of research, recommendations, and concluding remarks

The "shallow scrape" implemented here was completed relatively quickly for the six chemicals sampled at random. However, it can only provide counts of the various types of hits over calendar years. We used the shallow scrape methodology to investigate approaches to estimating the interest in harm caused by chemicals in the scientific literature. We considered two approaches to this task.

The first was to predict future concern based on historical publication records. The second was to detect a point in time at which the literature demonstrated a sharp increase in concern regarding some chemical, and thus, that this chemical was a candidate for further scrutiny by a regulator such as NSW EPA. These explorations caused us to identify particular areas worthy of further consideration.

Indicators of emerging concern and their use Our study sought to evaluate the usefulness of NSW EPA's previous approach of using mentions of chemicals in publications drawn from Taylor and Francis (T&F) publications. We determined that annual counts of T&F hits had a tendency to take lower values for individual chemicals compared to counts of hits obtained from the Web of Science database (WOS). We interpreted this as demonstrating that WOS had superior coverage of the scientific literature, and was thus a more appropriate database for our purposes.

We sought WOS mentions of our six sample chemicals for a range of query types. Of these, we concluded that the type seeking to determine those publication hits in which a chemical occurs within three words of some modifier relating to harm in the title or abstract ("near-hits" for each calendar year) was most appropriate. We made this judgement as we reasoned that near-hits would exclude those (extraneous) hits which did not relate to the chemical causing adverse effects on humans or the environment.

The use of near-hits allowed us to obtain reasonable predictions for the future near-hits our six chemicals. We interpret this as a validation of our approach. We also note that it may be possible to compare the results of predictions for a chemical with associated actual counts, and upon finding that reality greatly exceeded predictions, decide that the chemical warranted further scrutiny.

We recognised that an evaluation of the suitability of near-hits in detecting a sharp increase of concern from a historical record would be aided by a consideration of chemicals that were known to have provoked this response. As such, we drew upon the PFAS examples of PFOA and PFAS, and compared two alternative publication-count-based measures of concern. By considering the publication range from 1988 to 2018, we determined that for each of these PFAS, the time course of near-hits showed a distinct publication year after which there was a sustained increase. We found this signal to be sustained and unambiguous, and hence preferable to other measures of concern.

Given the nature of this pilot study, thus far we have considered a small number of chemicals. As such, whilst the use of near-hits as a means of quantifying emerging concern shows promise, we cannot make any claims as to its usefulness in general. As such, other measures may be more appropriate, as we noted in our discussion of hits which relate to particular health effects (Section 6.2).

Analysis of geographical locations and other keywords The "shallow scrape" of counts described earlier is unable to accessing keywords from publications. However, the implementation of a "deep scrape" able to obtain publication-specific information was frustrated by the data issues referred to in Section 3. Ongoing contact with Clarivate has led to some data cleaning, but owing to the time taken for this, most of CEBRA's efforts were directed by necessity towards the shallow scrape.

Modifying the queries outlined in Section 4.1 to suit the deep scrape, we searched WOS for individual hits for some chemical over the publication range of 2012-2018. Once all hits were obtained, we sought to extract publication information (e.g. associated journal title, keywords) from these individual publications. Obtaining the year of publication for each hit allowed us to form time series of the hits over publication years for any query, as we had for the shallow scrape.

The deep scrape for our six sample chemicals showed very few mentions of geographical locations in NSW. As such, should NSW EPA wish to pursue this investigation, it may be more helpful to begin by harvesting those WOS publications associated with a suitable location and some modifiers (such as "contaminant"), and to subsequently extract from these hits the mentions of particular chemicals over time.

By its nature, the deep scrape is substantially more time consuming than the shallow scrape. This is especially the case for high-prevalence chemicals which may have tens of thousands of associated publications in the "total-hits" case. Any further investigation of this option will certainly benefit from code optimisation. In a similar way, parallelisation of code will be particularly useful should NSW EPA wish to consider many more chemicals than the six studied here, especially if there is to be further experimentation with additional query types, and comparison of results.

Concluding remarks We note that at this point we have not yet answered certain questions, including "What is the best way to measure emerging concern?" However, we consider this project to be a successful initial exploration of the "technological space" of data sources, and methods for extracting information from them. We make this judgement for two reasons. The first is that results presented here demonstrate that it is possible for our methodology to recognise certain trends in data. The second is that for the PFAS examples we can discern a point the historical record at which the extent of attention given to chemicals in the scientific literature increases sharply— the type of event that one would associate with "emerging concern".

Given the familiarity gained with data sources and query conventions, and the body of R code developed thus far, in any future project CEBRA expects to readily incorporate information provided by NSW EPA into queries, and to productively scrutinise the results of these.

The next stage of the project will benefit from considering a larger number of chemicals. This will allow us to experiment with features such as other measures of prevalence, and more sophisticated models. We may then ascertain whether we can simi-

larly obtain accurate predictions for chemicals, or anticipate spikes in concern, as we demonstrated here.

Having accumulated a body of results, it will be appropriate to compare how particular chemicals are classified under the existing CPF and CEBRA's approach, and to determine the veracity of these classifications when compared against either publication data, or the actions of regulatory bodies.

Acknowledgements

This report is a product of the Centre of Excellence for Biosecurity Risk Analysis (CEBRA). In preparing this report, the authors acknowledge the financial and other support provided by the Australian Department of Agriculture, the New Zealand Ministry for Primary Industries, and the University of Melbourne.

Notices

A. Sample validation of CEBRA code

The results of a"Topic search" of WOS by web interface for a randomly selected chemical are shown in the screenshot of Figure A.1. We show "Topic Search" results for "pfna" OR "perfluorononanoic acid", for the publication year range 2014 to 2018.

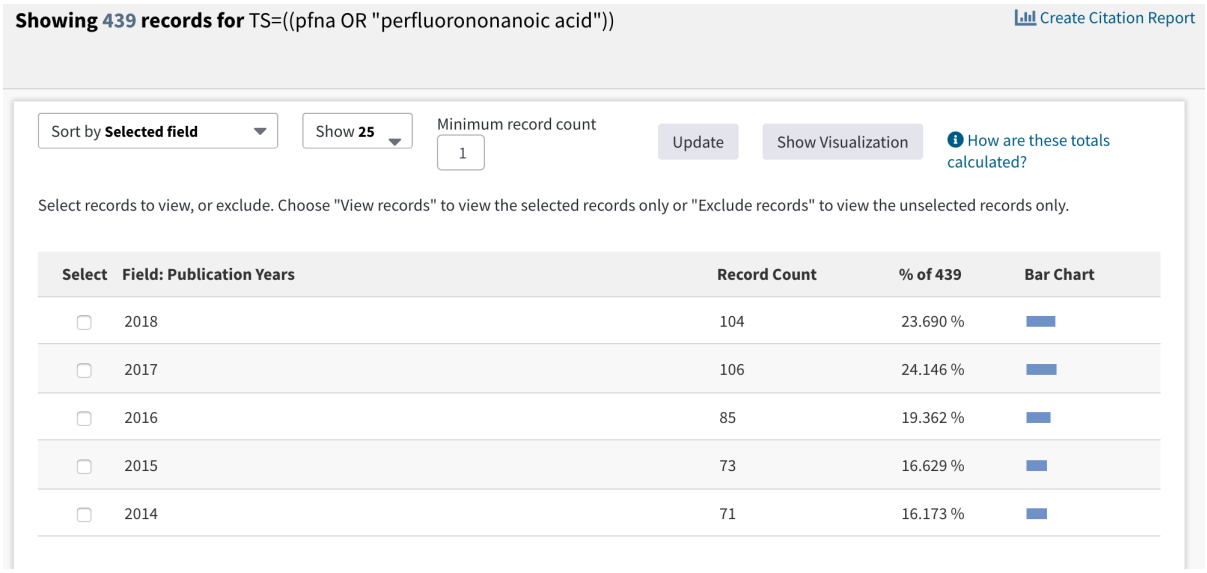

Figure A.1.: A screenshot of counts obtained from the WOS web interface for PFNA.

Counts obtained under the same conditions using CEBRA's R code are shown in Table A.1. Thus far we have only tested the total-hits, however we see that the counts decrease as expected as successive restrictions are added to the "total-hits" query. For brevity, the table only shows PFNA as the chemical name, but both of the alternative names were used in the search. This preliminary investigation shows that CEBRA's R code used to scrape WOS can reproduce the hits obtained for a randomly chosen chemical from the WOS web interface. Hence, CEBRA's code is operating reliably.

Table A.1.: Counts of hits for one chemical under the prescribed search conditions using CEBRA's R code applied to the WOS API. Note that the "total-hits" value for each year of the publication range agrees with the corresponding result shown in Figure A.1.

chemical	year range	total-hits	same-hits	near-hits
PFNA	2014	71	66	62
PFNA	2015	73	68	67
PFNA	2016	85	82	80
PFNA	2017	106	97	94
PFNA	2018	104	95	93