For what it is worth, here are a couple of thoughts I have about how to record, predict and then assess the accuracy of PEM classes.

1. Field recording of observed Site Series class(es).

I have said this before but here I just want to provide some rationale and examples.

I think that there are real advantages to recording the field observation data (training or accuracy assessment) as an estimate of likelihood of each possible class occurring at each observed position (or transect segment) as opposed to recording a single definitive class call or a first call and alternate call.

I say this because almost all locations in the field are intergrades between the centroids of idealised classes (e.g. prototypical location or central concept sites) and what is observed on the ground. There is no real “Truth” when it comes to conceptual classes.

A-Xing Zhu calls this Prototype Category Theory and sees all sites as having a Fuzzy Similarity to prototypes of all possible classes that could occur at any location (with 0 similarity for classes unlikely to occur at a site and 100 similarity if a prototypical individual occurs at a site). Alex McBratney and Florence Caree had a paper about assessing taxonomic distance or taxonomic similarity of a site to defined soil class taxa, which is essentially similar.

So, instead of asking your field staff to record a single definitive class for each site (or for each defined segment of a transect) you could instead provide them with a field form that would let them record their confidence in what class (or classes) occur at a site or along a segment. Think of these calls as "likelihood of this site matching the central concept of each of these N classes" OR "Confidence that I have that this site matches the central concept of these N classes" OR "Similarity of this site to the central concept of each of N possible classes in this area".

This kind of subjective assessment of likelihood could easily be accommodated in a field recording form that looks something like this:

BGC Variant:															
Name of Field Person:															
Date:															
Transect Number:															
Start Coordinates:															
Segment Number:	Baring	Start (m)	End (m)	01	02	03	04	05	06	07	08	WATER	ROCK	GRASS	OTHER
1	45	0	50	70	20	10	0	0	0	0	0	0	0	0	0
2	45	50	100	10	60	30	0	0	0	0	0	0	0	0	0
3	45	100	200	0	10	80	10	0	0	0	0	0	0	0	0
4	275	200	280	0	0	0	0	0	0	0	0	100	0	0	0
20	160	800	1000	0	0	0	0	40	50	10	0	0	0	0	0

In a form like this, the field ecologist would record their estimate of how likely the site (or transect segment) was to the central concept of each possible Site Series located in that BGC Variant or Subzone. Think of this as “How confident am I in my call as to what class best describes this site or segment?”. If the field ecologist is always 100% sure of the call for a single Site series, this form ends up having a 100 in one column and a 0 in all others for any given site or segment. So, this form can easily accommodate your current approach to collecting field data, as well as a more fuzzy approach.

2. Predicting Site Series class(es) using the field data.

Since your field observation data is now represented by an integer number from 0-100 that represents the likelihood of each class occurring at each site (or along each segment), You now have the ability to predict a numeric value for each class as the likelihood of each class occurring at each location. So, if you have 8 Site Series classes, you predict 8 maps of the possibility of a class occurring at each location. Possibility values range from 0 to 100 as per your input data.

One advantage of this approach is that you can record the absence of a class (0 likelihood) as well as the presence of a class (100 likelihood), so you can tell the model that some classes are virtually impossible to occur at a site, some may have a small likelihood and a few (1, 2, or 3) may have some significant likelihood of occurring at the site.

This is how most Machine Learning models work anyway. They want to predict the likelihood that each of N classes will occur at each location, then harden the classification by selecting the class with the greatest likelihood as the class to predict for that site. So, why not feed the model the data in the format that it prefers to use to make predictions, as probabilities or possibilities.

2. Assessing the accuracy of Site Series class predictions.

Since you have input data as integers on a scale of 0-100 and output maps as integers on a scale of 0-100 for likelihood of each class occurring at each location, you can use standard statistical measures to compare predictions to observations. You can use R2 or RMSE. You can compute the arithmetic difference between the predicted likelihood value for each class at each location and the observed or estimated value at field inspection locations. This means you are not assessing error in terms or a binary right-wrong metric but rather one that asks how similar or different is my prediction from my field observation. This makes for a much less stringent test and you end up with numbers that look a lot better, since most sites are not hugely different than the predicted site series, only slightly different. You capture and record these slight differences.

You can also assess the accuracy or correctness or the hardened map of single Site Series classes using a similar similarity approach.

In this approach, you harden the predictive map to predict a single, most likely class at each location. Then you compare this hardened class to the most likely observed class at each observation location. So, you compare a predicted 04 with an observed 05 or 03.You want to assess how good or bad this prediction is in terms of how similar the attributes of the predicted class are to the attributes of the observed class.

You can assess relative similarity using an approach based on a similarity matrix as illustrated below.

Option 2: Class similarity comparisons:

Example Similarity Matrix		01	02	03	04	05	06	07	08	WATER	ROCK	GRASS	OTHER
How similar is an 01 to an 02, 03	01	100	50	30	20	10	5	3	1	0	0	0	
and so on ...	02	50	100	40	30	20	10	5	2	0	0	0	
Can be calculated objectively	03	30	40	100	40	20	5	2	1	0	0	0	
based on moisture and nutrient	04	20	30	40	100	60	40	20	10	0	0	0	
ordinal ranking for each class	05	10	20	20	60	100	50	30	10	0	0	0	
The closer in moisture or nutrient	06	5	10	5	40	50	100	40	20	0	0	0	
class, the more similar the site series	07	3	5	2	20	30	40	100	40	0	0	0	
	08	1	2	1	10	10	20	40	100	0	0	0	
	WATER	0	0	0	0	0	10	20	50	100	0	0	
	ROCK	0	30	20	0	0	0	0	0	0	100	0	
	GRASS	0	0	29	0	0	0	0	0	0	0	100	
	OTHER												

If you predict an 02 and you observe an 03, this matrix would tell you that the predicted 02 is 40% similar to the observed 03. Similarly a predicted 02 would only be 2% similar to an observed 08.

I built similarity matrices like these as far back as 1984 when I was comparing soil series predicted classes to observed classes. I used a systematic and objective procedure to compute relative similarity based on the attributes used to define each class. Things like Drainage Class, Salinity Class, Texture Class, Calcareous class and others. For each of the defining classes, I had another similarity matrix that gave a number from 0-100 to relate, for example, similarity of well drained to moderately well drained to poorly drained. Then I calculated the similarity of each soil to every other named soil in terms of how similar they were in terms of their defined attributes of drainage, salinity, texture and so on. You could do that quite easily for Site Series, as most Site Series are differentiated mainly in terms of their position on moisture and nutrient gradients that are easily comparable numerically (at least ordinally).

One you have such a table that gives a numerical value for how similar any Site Series is to any other Site Series (in the same BEC Zone), then you can easily compare the similarity of a predicted class on the map with an observed class on the ground. The comparison is not right-wrong but rather a numeric value of how different. Most predictions are only slightly different than the observed class, so you get a much better accuracy number.

3. Some thoughts on old accuracy assessments I did in the 80s.

Gen asked me the following questions about some reports I authored on 2 projects I conducted in the mid 80's and early 90's to assess the relative accuracy of soil maps, irrigation suitability maps and various maps made using different approaches to manual soil mapping.

Gen's specific questions were:

Curious to know how the reports were received from your work in the 80's/90's – did you have any feedback about the proposed methods?

MacMillan, R. A., and R. M. Krzanowski. 1986. Evaluation of alternative ground truth procedures for soil mapping in Alberta. In: Proceedings of the 23rd Annual Alberta Soil Science Workshop. February 25 and 26, 1986, Calgary, Alberta. Alberta Soils Advisory Committee and Faculty of Extension, University of Alberta, Edmonton, AB. pp. 198-205.

Alberta Research Council., (ARC) 1992. Soil mapping systems: Part 1. literature review. by: R.A. MacMillan, W.L. Nikiforuk, M.D. Fawcett, R.L. McNeil and L.W. Turchenek. R.W. Howitt (ed.). Prepared by: Alberta Research Council, Environmental Research and Engineering Department. Prepared for: Agriculture Canada, Alberta Land Resources Unit, Centre for Land and Biological Resources Research, Edmonton, Alberta. funding provided by the Canada/Alberta Soil Conservation Agreement. 140 pp.

Alberta Research Council., (ARC) 1992. Soil mapping systems: Part 2. Options and rationale for new mapping systems. by: R.A. MacMillan, W.L. Nikiforuk, M.D. Fawcett and R.L. McNeil. R.W. Howitt (ed.). Prepared by: Alberta Research Council, Environmental Research and Engineering Department. Prepared for: Agriculture Canada, Alberta Land Resources Unit, Centre for Land and Biological Resources Research, Edmonton, Alberta. funding provided by the Canada/Alberta Soil Conservation Agreement. 187 pp.

Nikiforuk, W. L., M.D. Fawcett and R. A. MacMillan. 1993. An evaluation of the alternative methods of soil mapping. Environmental Research and Engineering Department, Alberta Research Council. Alberta Research Council Open File Report 1993-01. Edmonton, Alberta, Canada. 145 pp.

Fawcett, M. D., W. L. Nikiforuk, R. L. McNeil and R. A. MacMillan. 1993. An evaluation of the extrapolatory method of soil mapping. Environmental Research and Engineering Department, Alberta Research Council. Alberta Research Council Open File Report 1993-09. Edmonton, Alberta, Canada. 68 pp.

So, I can only provide very limited information on how these reports were received. The main reason I can not report much is because we did not seek, or receive, much feedback. There are explanations for this lack of feedback. I elaborate below.

First of all, we did not publish these studies in a refereed scientific publication. We produced them as limited circulation reports for the “clients” who commissioned them. In this case, the clients were ourselves (e.g. the Alberta Soil Survey and Alberta Agriculture - Irrigation Division). I believe that I can provide you with paper copies of these reports from my personal archives, should you be interested. They are otherwise not readily available. ***Not publishing in a refereed journal was a huge mistake and one I would urge you to avoid.*** If these types of studies are not published in a peer reviewed journal they do not receive adequate exposure, criticism or use. They have no persistence and have little impact on changing practices or ideas.

The second half of this equation was that we produced the reports for our own internal use and viewed ourselves as the “clients” for these reports. As such, we were not objective in how we evaluated or used the findings. We were biased in that we were actively seeking to “prove” that our maps were good and reliable and we massaged the results to achieve that interpretation. We had a vested interest in demonstrating that our maps were “as accurate as can be made” and we let ourselves convince ourselves that the data “proved” that. We did not honestly report on the limitations of our maps and did not report realistic levels of accuracy that could be attained for mapping hard classes in complex landscapes where all hard classes are simply one instance of realizations of a number of possible fuzzy classes. Does this sound familiar here?

As our own biased clients, we did a lot of mental gymnastics to convince ourselves that these maps were “good” and “came close” to predicting what was observed in the field at accuracy sampling locations. The powers that be then quietly buried these reports and announced that the results supported us to continue to make maps as we had always made them. I personally argued strongly for full and complete disclosure so that users would be able to judge for themselves whether the maps could address their needs. I even made it a point to publish results of accuracy sampling for Soil Survey project maps that I produced and published. But, honestly, there was a brief flurry of interest in statistical evaluation of map unit accuracy and soil property variability within map units within the larger soil survey community in a short period around 1979-1985 but then the interest all but disappeared. Senior management for soil survey in both Canada and the USA decided that these sorts of studies were eroding faith in conventional soil maps and soil mapping and so they actively campaigned to discourage further studies and suppress or hide the results from the few studies that had been completed. They were trying to protect conventional soil surveys from criticism and were trying to prevent funding sources from dropping funding for continuing soil mapping as we knew it. This obviously did not work. Soil surveys began to be shut down and discontinued from the late 80’s until there was virtually no active soil survey mapping in Canada by the mid 90’s. We might have had a better chance of survival if we had been more honest, more systematic and more statistically rigorous.

My take is that you can try to hide your warts but they will eventually be discovered and users will respond accordingly. Better to be honest, up front and scientifically rigorous and to put in place procedures for objectively and systematically assessing errors and working to find ways to continuously improve both your methods and your products. This applies just as much for PEM now as it did for soil survey in the late 80’s and early 90’s. You can’t really hide. So be up front and be rigorous and systematic in your efforts to improve continuously.

4. Some thoughts on spatial versus aspatial accuracy.

Gen and Will asked me why I was so confident in predicting that the current field accuracy studies would almost surely produce results that showed higher levels of accuracy for aspatial assessments of the proportions of classes within defined polygonal areas than for assessments that evaluated exact categorical match at exact spatial locations (exact correspondence of observed to predicted site series class at each specific grid cell location).

This was an easy call for me to make. It starts with my experience in reviewing the results of the duplicated field accuracy traverses that Dave Moon insisted were run for at least a few test transects undertaken for the Cariboo PEM mapping. As you know, Dave requested that 4 different field survey teams each traverse 4 different accuracy transects to produce independent assessments of the most likely site series at each location along each transect. The field surveyors did not know how the other survey teams had classified each transect nor did they know what site series had been predicted by the PEM map. Dave wanted to be able to gauge or estimate the measurement error that existed in the field traverses we were assuming would accurately reflect and capture "ground truth". The results were unexpected and revealing.

I do not have the results from this work at my fingertips but will try to find them and summarize them later. In general, however, we got really quite low levels of exact categorical match at exact spatial locations when the 4 independent traverses were overlaid on each other and compared. Exact categorical match ranged from lows of around 10% of total transect length to perhaps a maximum match of 70% for calls made by the 2 most experienced teams (who had very similar mental models of Site Series). There were no instances where calls made by all 4 field teams matched at exact spatial locations more than 30% of the time. So, my take on this is that it would be ludicrous to expect any map to achieve an exact categorical match at exact spatial locations along accuracy traverses at a rate that is greater than can be achieved by 4 local experts on the ground and in the field making these assessments with access to all possible relevant data and observations. No 4 independent observers even come close to agreeing at the 65% level of accuracy that is expected for a PEM map to achieve. So, how can we expect a PEM map to achieve that level of accuracy at exact locations?

So, why do calculations of accuracy improve for aspatial assessments of the proportions of each site series computed along the entire length or area covered by a transect? The answer appears to be related to the fact that differences and errors in classification of sites by field ecologists at exact spatial locations tend to average out over larger distances or areas. While their calls about absolute classes may differ in specific locations most ecologists tend to recognize about the same amounts, or proportions, of the main ecological classes within any examined area of some minimum size, such as a triangular accuracy transect.

Different field ecologists may differ at exact locations along a transect with respect to how to best classify a site. One recognizes a change from, for example 01 to 06 at a certain point but another does not recognize this change for another 20 or 30 meters, but then does. So their classifications do not exactly overlap or match along many individual portions of a transect when compared. Similarly the first ecologist may recognise a change from 06 back to 01 40-50 m before the second ecologist. It turns out that, on average, most ecologists tend to recognize similar proportions of Site Series along the total length of a transect but tend not to recognize the exact same Site Series at the exact same spatial locations. These differences in classification at exact spatial locations tend to average out over the total length of an accuracy traverse. So, if we calculate the average proportions of Site Series along each transect as assessed by different ecologists these tend to converge towards a common value that represents the mean value over some minimum sized area.

We can consider each location, site or grid cell along a transect to represent one possible instance or realization of all possible classifications or outcomes at that site. For most classifications, most sites reflect the fact that classifications are fuzzy and most sites can exhibit attributes that are associated with 2 or more possible classes. So, every site can exhibit a different value for its likelihood (or possibility) of belonging to ***n*** possible classes. Some sites may exhibit attributes that are similar to the mean or prototypical example of a class centroid. These may be easy to classify and may be recognized as belonging to the same class by multiple different independent observers. However, most sites are more ambiguous and do not fit the central concept of a class prototype easily. They exhibit attributes that could place them in 2 or even more possible classes. So, each of these sites may be classified as one class by one observer and a different class by a different observer, depending on the observer's experiences and biases.

If we consider every site or grid cell to be an instance of ***n*** possible realizations of ***n*** classes, we can better understand why and how different ecological field experts can end up classifying the same locations into different ecological classes. So, any accuracy field transect essentially represents one possible spatial realization of what classes exist at each spatial location. Another ecologist may interpret the field observations and produce a slightly different spatial arrangement of classes than the first ecologist. Neither is absolutely correct in representing “truth”. Both are equally valid realizations of the overall “truth” but with a different spatial pattern.

Similarly, any PEM map we produce using predictive methods represents just one possible realization of the outcomes of all possible classes in space for each location or grid cell. Neither the map or any field accuracy traverse actually represents absolute “truth”. Both represent one possible realization of possible classes, with hopefully the most highly likely class predicted at each spatial location but not necessarily so.

There is

5. Some thoughts on the need to identify and set targets or "standards" for PEM map accuracy.

I believe that the first thing you should do if you want to assess the accuracy of any classed ecological map is to establish objectively what level, or levels, of classification accuracy are theoretically and feasibly possible to achieve under real world conditions.

We need to first establish how consistently and how accurately currently defined ecological classes can be recognized in the field by experienced field ecologists actually standing at a site.

If we run field tests to objectively establish the levels at which 2 or more field ecologists can agree on the correct ecological classification at exact locations along a transect, we can then have objective data on which to base standards that we can realistically expect any predictive map to achieve. You cannot expect a predictive map to achieve a level of correspondence between predicted and observed class(es) that is better than what 2 or more field experts can achieve examining the same locations in the field.

Interestingly, if you were to have 2, 3 or 4 different field ecologists independently classify the same transect(s) you could then evaluate what is feasible to achieve in terms of not just exact categorical match at exact spatial locations but also all of the other comparison metrics that you might think of and want to set objective standards for. For instance, you can run calculations of degree of similarity or difference of classes recognized by 2 or more experts for the same locations to get objective measures of fuzzy similarity of classes, in cases where the exact same classes are not recognized by 2 or more field experts. You can also run all the same calculations that compute aspatial accuracy that we have previously run to assess PEM map accuracy. You can compare proportions of each class computed along the entire transit length as reported by each of ***n*** independent field experts. You can then compute mean overlap for ***N*** transets for each BGC subzone and mean accuracy (ovelap) by Site Series by BGC subzone.

Once you can objectively establish what levels of different types of accuracy (e.g. exact categorical match at exact spatial locations, similarity of classes at exact spatial locations, aspatial proportions of exact classes along entire transit lengths, aspatial proportions of both exact classes and similar classes (alternate calls) along entire transect lengths) you then actually have all the information you need in order to be able to set realistic goals or standards that any completed PEM map would be required to achieve. Any PEM map should be able to predict exact classes at exact locations, similar classes at exact locations and proportions of both exact and similar classes along entire transect lengths at about the same level as field ecologists can do when they classify the sites in the field. You would not expect a PEM map to be able to do any better than 2, 3, or 4 expert ecologists observing and classifying sites in the field. If a PEM map can achieve standards that are comparable to what 2, 3 or 4 trained ecologists can achieve when they examine and classify sites in the field along triangular transects then the map is doing as well at classifying unknown sites as if we were to send the ecologists out to observe and classify every unknown site (grid cell) in the area. At that point, the PEM map can be said to be as good as if you were to have every single unknown site visited and classified by an expert field ecologist.

Since we are using classifications obtained in the field by trained ecologists as the basis for both our training point data and the data used to assess the accuracy of any predictive map, we must first assess how much faith we can put in these field classifications. Currently, we call this “ground truth” and we assume that the ecologist making the classification calls in the field is always 100% correct and these in-field calls can be assumed to represent “ground truth”

Assessments of ecological class map error using field observations made by a local expert assume that the local expert is always correct and any predictive map is therefore incorrect in all cases of non-agreement.

For the Cariboo PEM, we undertook some efforts to evaluate the degree to which different local experts could agree upon the correct classification for exactly the same locations in the field. We considered this to provide an assessment of “measurement error”.

We had four different experts traverse and classify exactly the same four field transects at different times and with no opportunity to converse or discuss their respective assessments.

Some results from this comparison are listed in Table 1.

Table 1. Tabulation of results of an evaluation of measurement error among four local experts

Type of Agreement or Error Assessment	**Exact Class at Exact Locations Spatially Congruent Agreement**	**Proportions of Classes in Transect Compositional Agreement**
Average Primary Call Agreement (any 2 experts)	42%	64%
Minimum Primary Call Agreement (any 2 experts)	23%	37%
Maximum Primary Call Agreement (any 2 experts)	73%	86%
Average Primary Call Agreement (all 4 experts)	21%	Not Reported
Minimum Primary Call Agreement (all 4 experts)	11%	Not Reported
Maximum Primary Call Agreement (all 4 experts)	30%	Not Reported
Average Alternate Call Agreement (any 2 experts)	Not Reported	71%
Minimum Alternate Call Agreement (any 2 experts)	Not Reported	43%
Maximum Alternate Call Agreement (any 2 experts)	Not Reported	95%

Let us start by looking at the degree to which local ecological experts agreed with each other in the field by identifying the exact same class at the exact same spatial locations (e.g. when transect data were overlaid, the exact same class was recorded for spatially congruent points, or locations, along a transect). In the very best case, two experts agreed on the exact same classes at the exact same spatial locations 73% of the time, for one of the 4 transects. In the very worst case, two experts agreed at the same locations along one transect 23% of the time. On average, across the 4 test transects, any two experts agreed with each other at the exact same locations 43% of the time.

If we set a more strict expectation and calculate the degree to which all four experts agreed at the exact same locations we observe a low of 11% and a high of 30% agreement for any one transect and an average of 21% for all 4 experts for all 4 test transects. This pretty much confirms the old adage that if you ask 5 experts to all classify the same site you will get 6 different classifications. But this very low rate of agreement among 4 different experts really illustrates the fact that only a very small percentage (11-30%, mean 21%) of any area is so clear and unambiguous that all 4 experts see it and classify it identically. That means that the vast majority of any area (70-89%) has some ambiguity or fuzziness in its classification and is not assessed as identical by all 4 experts.

So, if we overlay an accuracy transect over any predictive PEM map, we should not expect that the map will agree with the field observed ecological classes at exact locations along the transect at a rate that is any greater than can be achieved by all 4 ecologists traversing the same transect length (minimum agreement 11%, maximum agreement 30% and average agreement across multiple transects of 21%). If we relax these criteria and set a goal of achieving levels of accuracy comparable to that which can be achieved between any 2 highly trained field ecologists, these numbers will then be (minimum agreement 23%, maximum agreement 73% and average agreement across multiple transects of 42%).

By carrying out these preliminary field studies, we can objectively determine what levels of exact classification accuracy we can feasibly hope to obtain for any predictive map of ecological classes. We should not expect to ever exceed exact class agreement at exact spatial locations on average at a rate of greater than 42%. And we should not be surprised if observed levels of exact class agreement at exact spatial locations are as low as 11-23% for any individual field transect. We can’t expect the predictive map to do better than can be achieved by 2-4 experts classifying the same locations in the field.

The accuracy metrics adopted for use in the Cariboo according to the Moon (2006) protocol were based on assessing overlap in estimates of proportions of ecological classes across the entire length of a transect compared to the proportions of classes estimated by the map for the equivalent spatial extent of the transect. We refer to this as aspatial accuracy or small area accuracy. The objective here is to achieve a minimum level in predicting the proportions of all possible ecological classes within a small area, defined as the area covered by a single triangular accuracy transect. Our triangular transects covered an area of approximately 17 ha, which was chosen to represent the minimum size plot area for which any management practice or activity might typically be planned and implemented.

The logic here is that, for many map uses, it is not essential to know what the classification is at an exact 25 m x 25 m pixel location. Most management applications are formulated for, and applied at, the scale of a 10-20 ha plot or cut block management unit. These management entities tend to be about 10-20 ha in size and decisions are taken for actions to be applied for the entire extent of the management plot. So, if we can estimate the proportions of ecological classes within these management sized areas with sufficient accuracy, then this map is likely to be of adequate quality to support the intended uses for making management decisions and plans at the level of a plot or block.

Often, we know that a misclassed site is actually very similar in most respects, to the correct class for that location. Consequently, it has been common to give partial credit for agreement when a field ecologist has indicated that they were uncertain of the correct class for a particular site or segment and indicated that there was a possibility that the site or segment might more appropriately belong to an alternate class, that is identified at the time the field traverse is run. With alternate calls, credit is given for agreement if the field ecologist indicated that a site might actually be best described by an alternate class and the map predicted that alternate class for that particular site or segment.

Local experts could not agree with each other on the correct proportion of classes along the tested traverses at more than 64% average agreement, 71% if credit was given for alternate calls. The highest level of overlap agreement between any two experts was 86% and the lowest was 37%. In terms of exact agreement of classifications at exact locations (spatially congruent accuracy) average agreement between any two experts was only 42% with the lowest exact agreement between any two experts being 23% and the highest exact agreement being 73%. All four experts agreed with each other at exact locations on average 21% of the time with a low of 11% exact agreement for one traverse and a high of 30% for the best. This small test appears to suggest that it would be unrealistic to expect any predictive map to agree with the proportions of classes observed by local experts along transects in the field at any better than 64-71%.

So, my proposal is that we start this project to assess PEM map accuracy by first running an initial project to assess levels of classification accuracy in the field that can be attained feasibly by local experts classifying the sites on the ground. We can assess classification accuracy at various different levels of categorical and spatial exactitude or generalization. Thus we can have the following measures of classification accuracy.

1. Exact categorical match at exact spatial locations
2. Exact categorical match or alternate call at exact spatial locations
3. Fuzzy similarity between classes observed at exact spatial locations
4. Aspatial assessment of overlap in proportions of exactly the same class calculated over the total length of a single field transect
5. Aspatial assessment of overlap pin the proportions of exactly the same class, OR a closely similar alternate call, over the total length of a single field transect
6. Aspatial assessment of fuzzy similarity of observed classes to predicted classes along the total length of a single field transect. (Or r2 or RMSE).

I suggest that you set a goal of having 1 in 20 triangular field transects repeated by having a second independent team follow the transect and classify it a second time independently.

Maybe you could have 1 in 50 run 3 times and 1 in 100 run 4 times. After you reach some minimum number of replicated transects (maybe 30?), I think you would have a pretty good handle on how accurately field ecologists could recognize and classify Site Series on the ground under different conditions and in different regions. These data could then form the basis for setting your expectations for what any PEM map should be able to achieve. If a PEM map can be as accurate as any 3 or 4 ecologists classifying in the field then I suggest that this map is as good as sending out ecologists to classify evey possible site (grid cell) in an area of interest.

I think that this approach to assess field classification accuracy before we try to assess PEM map accuracy is rigorous, systematic, objective and reproducible. It gives us a very sound basis for setting expected levels of field scale classification accuracy and, from that, of PEM map accuracy. We are not just grabbing numbers out of thin air. We have a sound basis for any numbers we select and advance.

And finally, I note that the initial numbers we got from 4 independent accuracy transects look suspiciously similar to the kinds of numbers you recently reported for all of the different types of exact spatial spatial and approximate aspatial accuracy that you got in your recent pilot area studies. If the PEM maps you assessed achieve accuracies similar to what we got from on the ground experts in all of these different measures then the PEM maps are likely as good as you can feasibly hope them to be.

I was enthusiastic about Dave Moon’s idea of running multiple repeats of classification along accuracy traverses to assess “measurement error” when he proposed this back in 2006. But I don’t think I fully appreciated all of the nuances and implications of this idea back then and maybe not even now. It is a really significant idea and it deserves to be taken seriously and implemented and analysed with rigour and dedication. This could be the basis for a very important and original study on classification accuracy (leading to applying the same methods and metrics for subsequent predictive map accuracy) that I have not seen anyone else ever do before. There is something potentially very exciting and important here.