Proposal of a synthesis paper based on FLUXNET database

Title: Paradigmatic networks: a new tool for gap-filling and spatialization of eddy covariance CO_2 fluxes.

Initial coordinators: Giovanni Manca[1], Stefano Federici[2],

Co-authors: Guenther Seufert[1], Giulio Marchi[1].

1. European Commission, DG Joint Research Centre Institute for Environment and Sustainability Climate Change Unit, Ispra, Italy.
2. Università degli Studi di Cagliari, Facoltà di Scienze della Formazione Dipartimento di Scienze Pedagogiche e Filosofiche

Short outline

The role of plant ecosystems in the global carbon balance is not fully understood and major questions remain open. In particular there is not a clear explanation of the CO_2 missing sink, namely the anthropogenic carbon dioxide emitted to the atmosphere each year that is not absorbed either by the atmosphere and the oceans. Recent researches show that northern mid-latitude terrestrial ecosystems and undisturbed tropical forests are important carbon sinks, but the magnitude, location and mechanisms are yet under investigation.

The FLUXNET database is an important source of information at local scale because it includes meteorological variables and carbon dioxide exchanges (NEE) measured at the interface between plant canopy and atmosphere using the eddy covariance technique. The experimental sites sharing data in the FLUXNET database represent the more important terrestrial ecosystems and this is an important feature needed for the spatialization of the site-level information. This source of data could be used for estimating the worldwide spatial distribution of carbon fluxes in order to study the role of different biomes in the terrestrial carbon balance. Obviously only through a spatial approach it will be possible to analyse and comprehend the inter-relationship amongst carbon fluxes and the environmental context. Therefore the scope of the spatial analysis will range from the local to the regional to the global scale in order to investigate the possible correlations between carbon fluxes and controlling environmental conditions such as land cover, climate or elevation.

Several global land cover classification products from remotely sensed data have been widely recognized and are available, e.g., the Global Land Cover Characteristics database (U.S. Geological Survey's (USGS), the Global Forest Resources Assessment (FRA 2000), the UMD Global Land Cover Classification (University of Maryland), the GLC2000 Global Land Cover (GVM-Global Vegetation Monitoring, Joint Research Centre of the European Commission). Global climate datasets are available from different sources, while elev ation data such as the Shuttle Radar Topography Mission (SRTM) are very often adopted for their high-resolution, extent and precision. The spatial approach will also allow the

interpolation of site-level data as well as the eventual zonal analysis with the relevant environmental parameters. Considering the somewhat course spatial resolution of 1km^2 of the land cover global datasets and the limited number of classes of these classification systems, the global approach may anyhow allow to compare homogeneously and effectively different sites for the same variables in similar environmental conditions and to derive informative trends and conclusions. At the regional / European scale, some tests will be performed on basis of more detailed data sources available at JRC, like the map of tree species/forest types or high resolution meteo data.

FLUXNET database could be used to test more sophisticated gap-filling methodologies. Gaps in the series of meteo data and CO_2 fluxes are a general problem affecting all sites. Different methodologies have been proposed: mean diurnal variation, non-linear regressions, look-up tables, simple process models, Kalman filter approaches and Monte Carlo techniques, like Multiple Inputation (MI). Some gap-filling methods have been already used for spatialization of carbon fluxes (e.g. neural networks).

We propose a new methodology for gap filling of meteorological and NEE time series, based on *paradigmatic networks*. This methodology will be used for worldwide spatialization of NEE values in order to get a map of annual carbon balance of terrestrial biosphere.

The new methodology proposed for this project is a learning technique, that is an inferential technique that hinges, as other learning techniques, on the general principle that good generalizations can be derived from observation of real data. To briefly summarize the main idea behind inferential techniques, they states that, for a given phenomenon, if condition A is the cause of a given result B, then by repeatedly observing the phenomenon with will see a certain noticeable amount of situations in which the given result B will be associated to (preceded by) condition A. A precondition for a learning technique being applicable to a domain is that the complexity of the association between A and B in the domain be only partially in the process that goes from A to B, mainly being in the (sometime huge) number of different situations (conditions A_1, A_2, A_3, …) that all go to B. That is, the solution to the problem must be envisaged by a simple algorithm with a lot of different preconditions. The ideal problem that is solvable by a learning technique should not be solvable by a complex algorithm based on a small set of preconditions.

Paradigmatic networks, even if capable of tuning their parameters exactly like other learning techniques and capable of reproducing the same phenomena they have previously learned, must not be confused with other well known learning techniques. Indeed, paradigmatic networks radically depart from statistical learning and other mechanisms such as neural networks or decision trees. The strongest feature of paradigmatic networks (and the most important difference with other learning mechanisms) is the emphasis that this mechanism puts on the relevance of *context*. Context can be defined as the set of *contour conditions.* This contour is exactly what the most part of learning mechanisms try to detect and to cut off, so that only “relevant” variables are taken into consideration. Except for very simple processes, this can’t be done without loosing a lot of relevant information about the process. A neural net or a statistical mechanism trying to learn several combinations of necessary/sufficient conditions could be misled to make the wrong inference. To give but a trivial example, let us suppose the following set of observational data

1) Adam is at home from 2 to 4 AND Brian stays at Adam's place from 1 to 3. THEN Brian is going to meet Adam.
2) Adam is at home from 2 to 4 AND Brian stays at Adam's place from 3 to 5. THEN Brian is going to meet Adam.

A generalization mechanism will tend to strengthen condition A (Adam is at home from 2 to 4) and to assign less importance to condition B_1 or B_2 (about Brian going to Adam's place). But, obviously, conditions B_1 or B_2 are necessary because condition A is not sufficient. That is, Brian has to go to Adam's place in order to meet him when Adam is at his place (condition C).

Why this is going to happen when applying neural networks or statistical learning to inferential tasks? The main reason is that the objective of these learning mechanisms is to "stack" observational data one onto the others, so that similarities are strengthened and differences made weaker. So, if several different conditions are responsible for a given result (e.g. CO2 flux) given a general stable set of contour (irrelevant or not sufficient) conditions, all the alternative (responsible) variables will tend to override each other's. The final set of supposedly relevant conditions will be given (for absurd) exactly by the set of irrelevant (or not sufficient) ones. This is the well known *overshooting effect* affecting neural nets and other "stack" learning mechanisms. Supporters of those mechanisms must be confident that in observational data (i.e. in real cases) this unfortunate co-occurrence of irrelevant/not sufficient stable conditions is not going to show up.

To avoid problems like this one, "stack" mechanisms have to carefully differentiate their input representations so that undesired similarities are not spotted. Instead, paradigmatic nets are completely unaffected by this kind of problems. The reason why this happens is that paradigmatic nets are a conservative mechanism: context is taken into the greatest consideration and contextual information is always kept, clearly linked to conditions that are common to several observational data. E.g., for the above example, the paradigmatic net would factorize the co-occurrence of condition A with conditions B_1 and B_2 so that, at the end of the inferential process, it will state that situation C will occur if and only if condition A AND (B1 OR B2) are met.

Paradigmatic nets have been used so far in a lot of inferential systems to analyze/produce natural language sentences and texts. Language is indeed one of the very interesting domains where the incredible number of variables at stake won't make succeed a mechanism that bases its inferences on a complex algorithm only based on a few cases.

Moreover, "stack" mechanisms are seriously affected by amount of repetitions. It can be the case that even if an observed case presents himself again as an input to the system, the system won't give the correct answer. This happens when a specific situation is in contrast to a general trend. As a consequence, only patterns belonging to several general trends can be learned by the system. Specific situations that do not conform to those general trends cannot be correctly answered. Again, this is not the case for a paradigmatic system. Specific cases are all answered correctly.

Finally, "stack" mechanisms such as neural nets have an artificial limit in the amount of observational data they can take into account (*saturation*). They cannot discriminate more than a maximum amount of different input patterns. Paradigmatic networks instead have their better performances when a large amount of observational data is fed to the net. More they learn, the better they behave. There is no degradation in their learning curve.

This is not the whole story about stack mechanisms and paradigmatic nets. Several other interesting differences hold, but one last thing must be noted. For a lot of different phenomena, very simple statistical mechanisms can give the right answer for a majority of the occurring situations (e.g. 80-90% of the cases). The problem arises when less general situations must be answered. Indeed, as specific contexts are discarded in favor of general trends, the mechanism gets unreliable in the remaining part of the cases (the last 10-20% of the cases). The results computed by the system in these cases could be instead extremely relevant to the global result (e.g. total amount of CO_2 fluxes, or local amounts of CO_2 fluxes).

To sum up, the main advantages of paradigmatic networks over neural nets and other statistical mechanisms are:

1. Paradigmatic networks emphasizes the relevance of *context,* so they are not affected by the *overshooting effect*.
2. In paradigmatic networks there is no need of differentiating their input representations so that undesired similarities are not spotted.
3. Paradigmatic nets can handle huge amounts of data and they are not affected by saturation.
4. Paradigmatic nets are also reliable on specific cases. All learned cases are indeed all correctly answered.

Data request:

According to our background and data availability, we would like to test the concept in a first step with data sets from European forest sites, then to apply to global forestry sites and to non-forest sites.

Co-authorship rules

- PIs and other scientists that shall give intellectual input are co-authors of the paper.
- PIs of the sites used in the paper are going to be co-authors. Moreover we are open to add a second co-author name for each site, if the PIs ask for.
- If the author's list will be too long, we will evaluate the possibility of a group-name explained somewhere in the text in accordance with editors and co-authors requests.