Degradation of chemicals in soil (Robin Oliver)

One of Syngenta's core activities is the invention and development of novel crop protection chemicals (herbicides, fungicides and insecticides). Syngenta test tens of thousands of chemicals in the hope that a few will become a new commercial Active Ingredient (AI). Testing can take up to 10 years and there are many chemicals that start this process that never go anywhere. There are many factors that affect whether or not a chemical is successful such as potency, selectivity, intellectual property and ease of manufacture but safety regulations require the compound to degrade in soil at a safe rate. Soil properties include the size distribution of particles, percentage of organic matter and microbial biomass and can be highly variable even within a single field. Moreover, the soil in which tests are performed is sieved which can also create an active environment different to the target environment. The soil degradation rates of test chemicals are measured approximately by Syngenta in a single soil at an early stage. Some chemicals are tested in more details (more soils, more replication, more timepoints, more precision of measurements). Any chemical that is put forward as a new commercial AI is tested much more rigorously using the regulatory standards. At each testing phase, the degradation rate is represented by a curve showing the change in concentration over time. At the early stage these curves will be very noisy.

Questions include: How good is the data being measured? How predictive are crude early tests of the outcomes of the definitive regulatory tests? What statistical models are appropriate here, e.g. functional data techniques could be used to create a model for a more accurate model for the degradation rate. A statistical approach does not take account of questions around the properties of the chemicals themselves and their interaction with the biological and soil environment, which would require a mathematical modelling approach, which in turn could lead to an inverse problem.

Coating of particles (Federica Cattani)

In a number of contexts, eg novel chemical catalysis approaches, Syngenta needs particles to be coated completely and evenly. Roughly speaking, this is done by adding a coating fluid to a box of particles and shaking the box. Another way of thinking of this is as a system of interacting particles that exchange 'information' on collision. Syngenta want to be able to describe these interactions analytically using abstract models, starting with pen and paper.

How can they be sure that the total particle surface area is covered by the fluid? What is the distribution of fluid on the particles? What is the optimum amount of fluid and shaking time? How long will a single particle remain without coating? Simulations can get so far but it is important to understand the distribution of properties and the stochastic dynamics of the particles. An important observation is that this may be related to the theory of so-called mixing times and cut-off phenomena of Markov chains in which an ordered system (fluid concentrated in one place) flips into a fully randomized state (fluid uniformly distributed randomly over particles) after a clearly identifiable number of steps (shakes).

Ranking new chemicals based on heterogeneous data (Russell Viner and Kostas Papachristos)

When seeking a new herbicidal AI, many chemicals are tested by applying them to multiple plants of multiple species at various doses and measuring the responses by eye (this is a herbicide "screen"). There are a number of different types of herbicide screen, which are usually run in sequence, and typically only chemicals showing activity in one screen will be tested in the second screen (people don't like to test "inactive" chemicals). This screening approach is good for selecting the best compound but bad for producing an overall ranking

of the compounds, as compounds are tested in batches and some are only tested in the initial screen, some have different dose rates, some are applied to different species and all batches have slightly different ambient conditions. The response activity of chemicals tend to correlate with the species of plant, and in this way a tree of responses can be built up, similar to a phylogenetic tree of the plant species.

The data collected for each chemical on each plant species is the percent control, ie 0% = the chemical did nothing, 100% = the chemical completely killed all plants. These data are used to fit a logistic regression model and an ED50 (effective dose, 50%) number is calculated. The ED50 is used to rank herbicides in terms of effectiveness on a given species.

Syngenta want to look for trends across a wider range of chemicals to aid design and ensure that the correct chemicals are chosen for further work. Is it possible to find an overall ranking of chemicals using many sets of pairwise comparisons in a large database? Can effects associated with the species of plants being tested or with the assessor who evaluates the effectiveness be taken into account. More generally, how can Syngenta model and thus explore the infinite-dimensional chemical space in a more efficient manner using e.g. unsupervised learning to group herbicides in terms of effectiveness.

Formulation toxicity (Kim Travis)

A complete formulation contains one or more active ingredients (AIs) and a number of other chemicals which are collectively known as co-formulants (solvents, surfactants, anti-foaming agents, anti-corrosive agents, sunscreen etc.) to create the product. The total list of all co-formulants used in all formulations is fairly long, but many are used in a lot of different formulations. In the process of designing a new formulation, the toxicity of these final formulations need to be predicted, given what is known about their component parts. Although the toxicity of all AIs individually will be known, for many of the co-formulants their toxicity is unknown (but similar chemicals tend to have similar toxicity). Currently, Syngenta assume that ingredients act additively. Can we use the data to show that specific combinations matter? Can we analyse the database of many formulations to obtain information about the toxicity of individual co-formulants (inverse problem). Syngenta have preliminary data but the nature of the information makes it difficult to analyse the whole effectively, as some variables are continuous and quantitative and some are discrete and qualitative. A route to solving the problem is required. This could have statistical elements to it but also requires expertise from numerical analysts.

If the available data are organized in a matrix, where each row corresponds to a component and each column corresponds to a formulation, then this matrix is sparse, as there will be only a few nonzero entries. Since the data are incomplete, one could try to “learn” the missing information by using matrix completion strategies, with some underlying assumptions. For instance, if similar compounds have similar behavior, then one might ask for a matrix of low-rank. This procedure is reminiscent of the “Netflix Problem”, where the goal is to predict user ratings for films, given some available ratings of some films.

The problem can also be seen as a classical statistics problem where the toxicity of each component are the unknown parameters, the proportion of each component in the formulation are the predictors, and the formulation toxicity is the response. If the ingredients act additively, then a multiple regression or spline model can be used.

Droplet analysis (Gordon Bell)
Spray effectiveness and droplet impact. A better understanding of spray coverage on leaves is an important problem; this requires for example, maximising the amount of product that remains on leaves and minimising what reaches the ground. Also maximising the evenness of coverage. There are many aspects to this problem: spray droplets carried by (possibly turbulent) air currents; impacts of drops on solid surfaces with possibly complex wetting characteristics; drop coalescence and flow; and the effect of the geometry of leaves and the configuration of leaves on a plant.

Plated cell experiments (Kim Travis)
Syngenta are interested in gaining a better understanding of developmental defects and the potential for their causation by certain chemicals. It transpires that Syngenta do some relatively fundamental-level research on the effects of chemicals on cells in petri-dishes (ie in vivo experiments). It is also possible to model these sorts of experiments using volume-exclusion processes and to do parameter inference (using approximate Bayesian computation perhaps) in order to determine the effect of chemicals on the migration, proliferation, adhesion properties of cells in comparison to untreated cells. This might shed light on potential developmental defects caused by these chemicals, which are often a result of impaired migration, proliferation or altered cell-cell interactions.

Dermal absorption (Kim Travis and Gordon Bell)
Deterministic mathematical models could be used to explore the impact of physico-chemical properties on the potential for a pesticide formulation to penetrate through the skin. For single chemicals, eg AIs, this has been well studied. However, understanding the influence of a co-formulant on the dermal absorption of an AI is a major unsolved challenge. Reaction terms such as binding and partitioning have the potential to affect the spatial movement of chemicals and can be easily incorporated qualitatively into spatial models [1]. The flux of molecules through the skin may depend simply on passive diffusion, or there may be some type of enhanced diffusion whereby the formulation facilitates movement by widening channels through the skin. One big challenge will be to determine boundary conditions at the skin surface and its lower boundary. However, the modelling process could elucidate the impact of these conditions on the amount of AI successfully delivered across the skin.

[1] Jones, White, Delgado-Charro (2016) Mathematical biosciences 10.1016/j.mbs.2016.08.007

Crop production scheduling (Kim Travis)
With twenty-year database of weather and yield data, is it possible to simulate and predict based around different scheduling strategies for land use when growing crops. This is a spatio-temporal optimization problem similar in spirit to scheduling in e.g. the NHS. At the moment there are no realistic means in place for performing an optimization. Forecasting needs to take account of the risk involved as well as the financial returns.