生物データの取り扱い方

担当：遠藤・植山
実習場所：B11-118 (情報処理演習室)

１．実習の目的

本実習では、実験実習や卒業研究などで取得される科学データの持つ意味を正確にわかりやすく伝える技術を身につける。科学データの整理方法を学び、わかりやすく伝えるための図表の書き方について学ぶ。

２．使用データと表計算ソフト

使用するデータは、国内の森林において閉鎖型チャンバー法（図 1）を用いて観測された森林土壌の呼吸速度とその場所の地温である。それぞれのデータは 30 分毎に取得されたものであるが、今回の実習では 1 日平均したものが配布される。データの 1 列目には日にち、2 列目には地温、3 列目には土壌呼吸速度が保存されている。また、1 行目にはデータの種類、2 行目にはデータの単位がヘッダー情報として書かれており、実際のデータは 3 行目以降から始まる。なお、正しく測定ができなかった日については、“-9999”が異常値として記入されている。

データ整理には、表計算ソフト(Excel 2013, Microsoft®) を用いる。表計算ソフトは、平均や標準偏差等の統計値を計算することができるほか、図表としてデータを視覚的に整理することができる。

図 1 森林土壌に取り付けられた土壌呼吸測定用チャンバー

使用するデータは、生態気象学研究グループのホームページ（http://www.envi.osakafu-u.ac.jp/atmenv/）にアップロードされている。

３．Excel を用いた作図

3-1. 散布図を用いた時系列グラフ

土壌呼吸速度と地温の季節変化を散布図で示す（図 2）。散布図とは、縦軸と横軸の値を対応させた点を示した図である。今回は、横軸を日（時間）、縦軸を土壌呼吸速度と地温とすることで、土壌呼吸速度や地温の日々の変化を図示する。

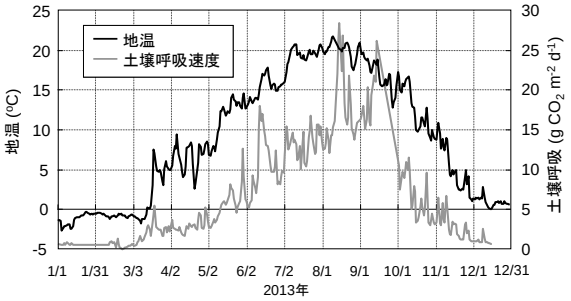


図 2. 森林土壌で計測された地温と土壌呼吸速度の季節変化

■Excel 2013 による作図方法

1. 異常値を除いた作図のためのダミー列の挿入

Excel では、異常値（-9999）を除いて作図することができない。ただし、「#N/A」値であれば、異常値を読み飛ばして作図される。そこで、IF 文を使用して異常値を「#N/A」値に置き換えた列を作成する。

例： =IF(C3 = -9999, NA(), C3)

2. 散布図を書く

Excel で標準出力されるグラフは、汚く見るに堪えない質であるため、見栄えよく調整する必要がある。ここでは、図 2 を目指すべきグラフとして、それに対して汚いグラフの例を図 3 に示す。図 3 で調整すべき点は、以下の通りである。

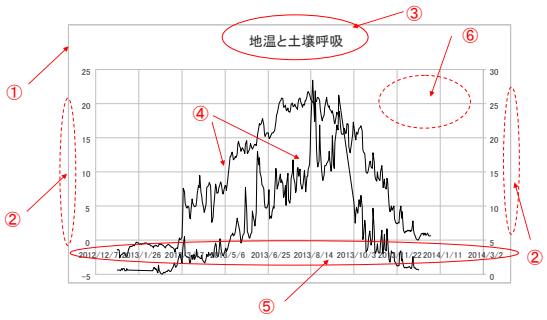


図 3. Excel 標準出力の汚いグラフの一例

① 図の枠線は消す。

② 軸見出しをつける。例えば、「地温 （℃）」や「土壌呼吸速度 （g CO_2 m^{-2} d^{-1}）」

- CO_2 の「2」は、必ず下付き文字とすること。
- 単位にある-2 乗や-1 乗は、必ず上付き文字とすること。

③ 図のタイトルは、下に付けること。また、図表のタイトルはそれを読んだだけで何を示しているかが理解できる程度に詳しく、かつ簡潔に記載すること。

④ どの線が何を表しているかがわかるように、濃淡、種類（実線、破線、一点鎖線など）、太さ、色などを調整すること。

⑤ 目盛のタイプや区切りを標準のままにしないこと。この場合は、「月/日」がわかれば十分で年の情報は不要である。

⑥ 線の凡例（Legend）を付けること。

その他、図 3 の例には無いが、軸目盛の有効桁数を合わせることなど、注意が必要である。

使用する Excel 関数

IF 関数 ： =IF(条件, 条件が成り立つ場合の式あるいは値, 条件が成り立たなかった時の式あるいは値)

例文 =IF(C3=-9999,NA(),C3)

NA 関数 ： =NA()

異常値（#N/A）を代入する関数 （N/A は、Not Available の省略語）

3-2. 回帰分析

土壌呼吸速度は微生物や植物の根の呼吸の総和である。一般に、植物や動物の呼吸速度は温度が上昇するとともに増加するとされている。このような要素間の関係性を定量化する手段として回帰分析が用いられる。回帰分析は、変数 X と変数 Y との間にどのような関係が成り立つか、あるいは関係が成り立たないかを評価する際に用いられる。変数 X と Y の間に、線形の関係（例えば、$Y = a \times X + b$）がある場合は線形回帰、指数の関係（例えば、$Y = a \times \exp(X \times b)$ ）がある場合は指数回帰を用いる。ここで、回帰分析によって得られる式は回帰式、回帰式の係数は回帰係数と呼ばれる。回帰式の当てはまりを表す指標として決定係数(R^2)や相関係数(R)といったものが用いられる。相関係数は、-1.0～1.0 の値をとり、絶対値が 1.0 に近ければ、当てはまりが良い、あるいは変数 X と Y の間の関係性が高いことを表す（図 4）。一般に、相関係数の絶対値が 0.7 を上回ると、両変数間に強い相関があるとされる（Wikipedia 「相関係数」, 2014）。

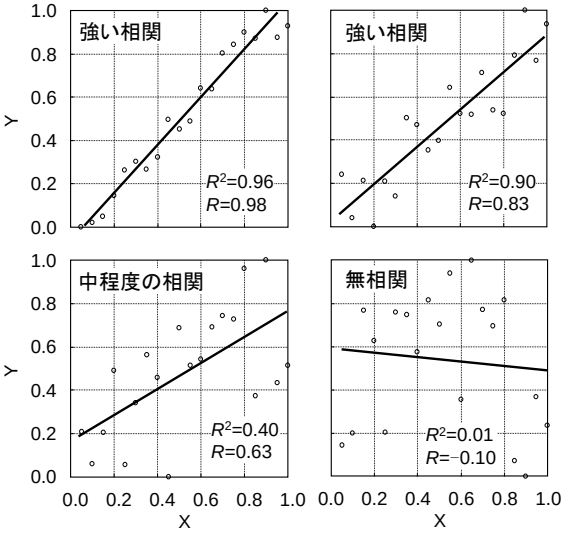


図 4. データの分布と相関係数に関する散布図

■Excel 2013 による相関解析

1. 地温と土壌呼吸速度を選択して散布図を書く（図 5）。

土壌呼吸速度と地温の間の関係を考える場合、地温の変化によって土壌呼吸速度が変化すると仮説が立てられる。逆に、土壌呼吸速度の変化で地温が変化するとは考えにくい。つまり、土壌呼吸速度は地温によって説明されるといえる。このような関係性が成立する場合、土壌呼吸速度のことを目的変数（従属変数）、地温のことを説明変数（独立変数）という。属性がはっきりしている両変数間の散布図を書く場合、横軸を説明変数、縦軸を目的変数としてデータをプロットする決まりがある。

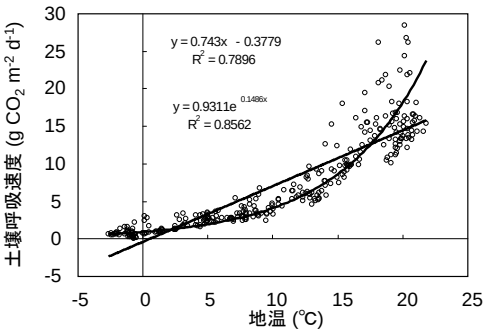


図 5. 地温に対する土壌呼吸速度の関係

2. 散布図のオプションを指定して、近似線を追加する。

今回は、線形回帰式と指数回帰式の 2 つを図中に追加し、いずれの決定係数が大きいかで関数の当てはまりの良さを調べる。

図 6 は、Excel による汚いグラフの一例である。既に図 3 で指摘されている点に加えて以下の点に注意して作図すること。

⑦ 縦軸の横目盛は、左側に詰める。横軸については、下側に詰める。
⑧ 回帰式の数字がグラフのプロットに重なっている。
⑨ 回帰式は太く見やすく書くこと。

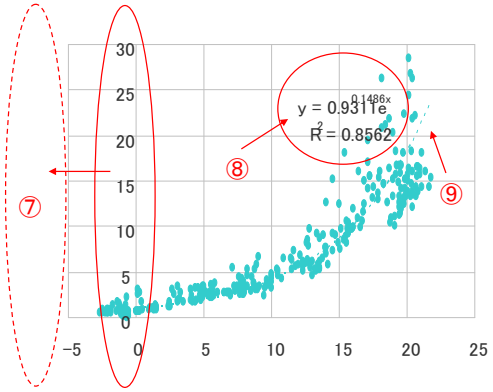


図 6. 汚いグラフの例

3-3. Bin データ

回帰分析において、目的変数や説明変数が大きくばらつくために明確な関係性が評価できない場合がある。これは、データの信頼性が低くそれぞれのデータの不確実性が大きい場合、2 変数以外の要因によってデータがばらついている場合（例えば、今回であれば土壌含水率などの他の要因によって土壌呼吸速度が変化している場合、地温と土壌呼吸速度との関係が不明瞭となる）などが挙げられる。このような場合、説明変数の値が近いいくつかのデータを平均して、平均された両変数間の相関解析を行うと関係性が明瞭になることが多い。このように幾つかのデータをまとめたデータは Bin データと呼ばれる。

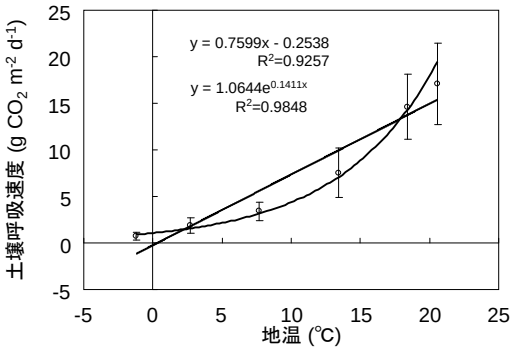


図 7. 地温に対する土壌呼吸速度の関係。データは 5℃毎の地温のクラスにまとめられている。

■Excel 2013 による Bin データの作成　（図 7）

1. Bin データを用いた回帰分析

地温が 5℃毎のクラス（例えば、0～5℃、5～10℃、10～15℃）で地温と土壌呼吸速度の平均値、標準偏差を計算する。

1-1. Bin データを作成するために、Excel 中の地温、土壌呼吸速度の行を別の行にコピーする。
1-2. コピーした行を選択して、地温についてデータの並び替えを昇順で行う。
1-3. 地温のクラス（例えば、0～5℃、5～10℃、10～15℃）ごとにデータセルを色分けする。
1-4. 同じ色のセルを選択して、地温、土壌呼吸速度のそれぞれについて平均、標準偏差を計算する。

2. Bin データの作成

上記の処理で作成された Bin データについて、散布図を書いて線形回帰、指数回帰分析を行う。ここで、各散布図のプロットに標準偏差を誤差線として書き加える。

使用する Excel 関数

Average 関数　　：　=Average (複数のセル; 例えば、A1:A100)

選択したセルの平均値を計算する。上例であれば、A 列の 1~100 行目までの平均を計算する。

Stdev 関数　　：　=Stdev (複数のセル; 例えば、A1:A100)

選択したセルの標準偏差を計算する。上例であれば、A 列の 1~100 行目までの標準偏差を計算する。

標準偏差とは、データのばらつきの程度を表す指標である。

3-4. 回帰モデルの精度評価

回帰分析によって得られた線形回帰式や指数回帰式の当てはまりの良さを調べる。ここで、回帰式から予測される変数値（この場合は、地温から回帰式により予測される土壌呼吸速度の値）のことを推定値と呼ぶ。目的変数と推定値との比較から得られるさまざまな統計値を用いて当てはまりの良さを知ることができるが、ここでは線形回帰式から得られる傾き（$Y = a \times X + b$ の a）、切片（$Y = a \times X + b$ の b）、相関係数（R）について述べる。

目的変数（観測された土壌呼吸速度; F_{obs}）と推定値（推定された土壌呼吸速度; F_{model}）から以下の線形回帰式を得た場合（図 8）、

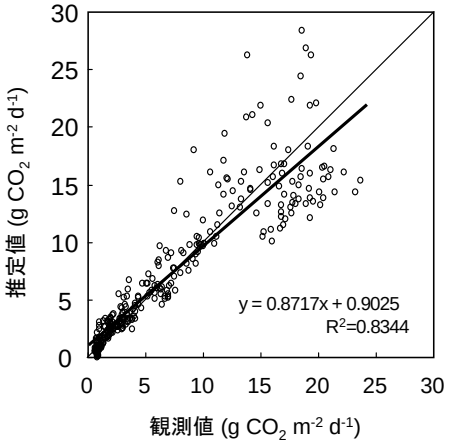


図 8. 指数回帰式による推定値と観測値による土壌呼吸速度の比較。細線は 1:1 線、太線は両者の直線回帰式を表す。

$$F_{model} = \text{slope} \times F_{obs} + \text{intercept} \qquad (1)$$

回帰係数 slope が 1.0 に近く、回帰係数 intercept が 0.0 に近い場合、回帰式の当てはまりが良いといえる。一方、intercept が正の値に大きくなる、或いは slope が 1.0 よりも大きくなれば推定値が過大評価されていることを表す。逆に intercept が負の値をとる、或いは slope が 1.0 よりも小さくなれば推定値が過小評価されていることを表す。

■Excel 2013 による回帰モデルの精度評価

1. モデル値の計算

3-2 節、3-3 節で推定された線形回帰式、指数回帰式を使って土壌呼吸速度の推定値を 4 つ計算する。

2. 線形回帰分析による推定値の評価

各推定値の当てはまりの良さを、3～8 月の期間について調べる。Excel では、散布図の近似線の回帰係数を調べる関数（SLOPE、INTERCEPT）や相関係数を計算する関数（CORREL）があり、これらの関数を用いて当てはまりの良さを調べる。Excel で、これらの関数を使用する場合、解析対象セルに

「#N/A」値が入っていると結果が「#N/A」となってしまう。そこで、「#N/A」値を空白に置き換えておく作業が必要となるが、今回のデータでは 3〜8 月の期間に「#N/A」値が入っていないので問題とならない。「#N/A」値を空白で置き換えたい場合は、3-1 節で作成した異常値を除去する下記の式を

例文　=IF(C3=-9999,NA(),C3)

次のように書き換えるとよい。

例文　=IF(C3=-9999,””,C3)

今回の推定値の評価においては、1 式にならって横軸（x 軸）を F_{obs}、縦軸（y 軸）を F_{model} として Slope と intercept、そして決定係数（相関係数の 2 乗）を計算する。

使用する Excel 関数

EXP 関数　：　=EXP(a)
　e を底とする数値 a のべき乗を計算する。
Slope 関数　：　=Slope (既知の配列 *y*, 既知の配列 *x*)
　$y = a \times x + b$ の a を計算する。
Intercept 関数　：　=Intercept (既知の配列 *y*, 既知の配列 *x*)
　$y = a \times x + b$ の b を計算する。
Correl 関数　：　=Correl (配列 a, 配列 b)
　配列 a と配列 b の間の相関係数を計算する。
Power 関数　：　=Power (数値 a, 指数 b)
　a^b を計算する。

Excel の相対参照と絶対参照

Excel では、セルの参照形態に相対参照と絶対参照の 2 種類がある。例えば、以下の式は相対参照の式であり、A 列 1 行目と B 列 1 行目の積を計算するものである。

=A1 * B1

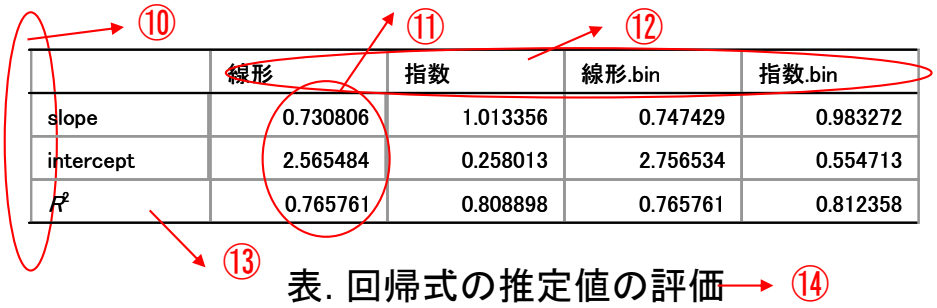
この式が C1 に書かれている場合、式を C2 にコピーすれば「=A2 * B2」となり、C3 にコピーすると「=A3 * B3」となる。また、D2 にコピーすれば「=B2 * C2」となり、参照するセルが相対的に移動する。一方、A セルを絶対参照で記載した以下の式の場合、

=$A1 * B1

表 1. 観測された土壌呼吸速度に対する回帰式による推定値の関係。Slope、intercept、R^2 は線形回帰式から算出された。1.0 より大きな slope、0.0 よりも小さな intercept は、推定値の過大評価を示す。

	線形回帰	指数回帰	線形回帰.bin	指数回帰.bin
slope	0.73	1.01	0.75	0.98
intercept	2.57	0.26	2.76	0.55
R^2	0.77	0.81	0.77	0.81

表 2. 推奨される書き方が表 1 である場合の作表の悪例。



式を D2 にコピーすれば「=$A2 * C2」となり、A 列に対する位置は変わらない。あるいは、

=$A$1 * B1

と記載すると、D2 にコピーしても「=$A$1 * C2」となり、A1 セルへの参照位置は変わらない。このように Excel では、セルを参照する場合に「$」を付けると絶対参照となる。「$」を付ける位置としては、「$A$1」、「A$1」、「$A1」の 3 つのパターンがあり、それぞれ「行列の固定」、「行の固定」、「列の固定」を表す。「$」は「F4」キーを用いると簡単に挿入できる。例えば数式中の A1 を選択して「F4」キーを押すごとに、$A$1（行列の固定） → A$1（行の固定） → $A1（列の固定） → A1（相対参照）の順に切り替わる。

表の書き方

モデルの当てはまりを評価するために 3〜8 月の期間の slope、intercept、R^2 を表 1 のような体裁で作図せよ。表 2 にしめされる悪例とならないように気を付ける作表する。

⑩ 表に縦線は不要。
⑪ 無駄に小数点以下の桁数を示しすぎない。有効桁数を意識して作表する。
⑫ 表の項目は分かりやすい名前を付ける。
⑬ 表中のフォントは、明朝体系のフォントを使用すること。「Times New Roman」などのフォントが推奨される。

⑭ 抽象的な表タイトルを使用せず、タイトルから表の中身をわかるような具体的なタイトルとする。また、表のタイトルは、表の上に書くこと。

４．レポート

1. 実習で取り扱った地温と土壌呼吸速度のデータを使って以下の図表を作成せよ。作図の際は、①～⑭で指摘された点に注意して作図すること。

   地温と土壌呼吸速度の季節変化を示す散布図 (図 2)
   地温に対する土壌呼吸速度の関係を示す散布図 (図 5)
   （線形回帰式、指数回帰式、それぞれの決定係数を図表に含めること）
   地温に対する土壌呼吸速度の関係を示す Bin データを用いた散布図 (図 6)
   （線形回帰式、指数回帰式、それぞれの決定係数を図表に含めること）
   回帰式の推定値の当てはまりを示す slope、intercept、決定係数を表す表 (表 1)
   線形回帰、指数回帰、Bin データを用いた両回帰式による推定値の評価を行う。
   評価の期間は、3-8 月についてとする。

2. 別途配布する森林における気温と蒸発散速度のデータについても、上記の地温と土壌呼吸速度のデータと同様の図表を作成せよ。ここで、気温の Bin データを作る際は、5℃毎（例えば、0～5℃、5～10℃、10～15℃）にクラス分けすること。また、推定値の当てはまりに関する表は、4～8 月の期間について示せ。

3. レポートは上記の図表を印刷の上、本日から 2 週間以内に植山（B11 棟 239 号室）まで提出すること（私が居室している時間であれば最終提出時間は問わない）。図表は、課題 1、課題 2 についてそれぞれ 1 ページでまとめること。遅れたものは採点しないので、注意すること。