Speciale i Matematik

Modellering af bakteriesammensætningen i et biogasanlæg

Af Babyasha Sritharan
Afsluttet den 10. Januar 2019

AALBORG UNIVERSITET
STUDENTERRAPPORT
Institut for Matematiske Fag • Skjernvej 4A • 9220 Aalborg Øst • Tfl.: 99 40 99 40

AALBORG UNIVERSITET
STUDENTERRAPPORT

Institut for Matematiske Fag
Matematik
Skjernvej 4A, 9220 Aalborg Øst
Tfl.: 99 40 99 40

Titel:

Modeling the bacteria composition in a biogas system

Modellering af bakteriesammensætningen i et biobrændselsanlæg

Projekttype:
Speciale

Projektperiode:
3. september 2018 - 10. januar 2019

Projektforfatter:

Babyasha Sritharan

Projektvejleder:
Lektor Poul Svante Eriksen
for Institut for Matematiske Fag
ved Aalborg Universitet

Sidetal: 109
Afsluttet den: 10. januar 2019

Synopsis:

I specialet estimeres en tidsserie over abundansen af OTU'erne (enhed for en mikroorganisme) i et biogasanlæg. Dette vil bidrage til ny viden indenfor biogas-teknologien. Der arbejdes med et datasæt, som indeholder oplysninger omkring en prøveserie udtaget i et af Maarbjergs biogasanlæg. Prøveserien er taget tre forskellige steder i biogasanlægget henover 27 ikke-ækvidistante tidspunkter. Udtagningen af prøveserien er i begyndelsen blevet udført ved at tage triplikater, hvorefter dette blev afbrudt grundet økonomiske årsager.
For hver triplikat blev der udvalgt en repræsentativ prøve. Dette blev blandt andet gjort ved at visualisere heatmaps og PCA biplot. Spline anvendes til at interpolere ækvidistante tidspunkter. Herefter estimeres en sparse VAR(1). Lassostraffen blev bestemt ved krydsvalidering. Baseret på strukturen i VAR(1) estimeres en VARX(1). Estimaterne i VARX(1) anvendes til estimering af Lotka-Volterra modellen. Ud fra de estimerede modeller kan det konkluderes, at dynamikken mellem mikroorganismerne er meget kompleks.

Forord

Dette speciale er skrevet som en afsluttende del af min kandidatuddannelse i matematik på Institut for Matematiske Fag på Aalborg Universitet i perioden 03.09.18-10.01.19. Med inspiration fra NomiGas-projektet [1] vil jeg under mit speciale beskrive den tidsafhængige dynamik mellem mikroorganismer i et biogasanlæg.

Jeg vil gerne takke min vejleder, Lektor Svante Poul Eriksen, for at have vejledt mig fra ide til slutfasen. Især diskussionerne med Svante har styrket min forståelse for statistisk behandling af biologisk data. Ydermere vil jeg takke Professer Jeppe Lund Nielsen, for at have inspireret mig til NomiGas projektet, Postdoc Nadieh De Jonge for udlevering af datasættet, og Caroline Kragelund Rickers fra Teknologisk Institut for at have drøftet fremtidige muligheder for at implementere tolkninger af mikrobielle data. Informationerne omkring datasættet er fået igennem dialog med Nadieh De Jonge.

Slutteligt vil jeg takke min mor Thevarani Sritharan for at have givet mig støtte igennem hele min uddannelse.

Abstract

In the thesis the abundans of OTU's (unit for a microorganism) from a biogas plant has been modelled as a timesseries. This will contribute to new knowledge within the biogas technology. The data used in the thesis contains information about a sample series, which has been taken in one of Maarbjerg's biogas plants. The sample series has been collected in the biogas plant at three different places over 27 non-equidistant times. The sampling series was initially performed by taking triplicates, after which it was interrupted due to economic reasons. A representative sample was selected for each triplicate, which was done by visualizing heatmaps and PCA biplot. To interpolate equidistant times spline was used. Then a sparse VAR(1) was estimated. The penalty term in the Lasso regression was determined by cross-validation. A VARX(1) was estimated based on the structure of sparse VAR(1). Finally, a Lotka-Volterra model was estimated based on the estimates from VARX(1). It can be concluded that the dynamics between the microorganisms are very complex, due to the estimated models.

Indholdsfortegnelse

Kapitel 1 Indledning **1**
1.1 Principal komponent analyse . . . 3
1.2 Spline . . . 6
1.3 Lasso . . . 9
1.4 Tidsrækkeanalyse . . . 12
1.5 Lotka volterra model . . . 14
1.6 Modelvalidering . . . 15

Kapitel 2 Metode **18**
2.1 Databeskrivelse . . . 18
2.2 Validering af sekvenserne . . . 19
2.3 Udvælgelse af OTU'er til modellering . . . 19
2.4 Interpolation af ækvivalente tidspunkter . . . 21
2.5 Modellering . . . 21
2.6 Modelvalidering . . . 21

Kapitel 3 Resultater **23**

Kapitel 4 Konklusion **31**

Bibliografi **32**

Appendiks A R-koder **34**
A.1 Eksempler . . . 34
A.2 Uoverensstemmelser . . . 35
A.3 Validation af sekvenser . . . 36
A.4 PCA, heatmap og spline . . . 39
A.5 Modellering af Lotka Volterra . . . 43

Appendiks B Heatmaps **44**

Appendiks C Splines **50**

Appendiks D Funktioner **62**

Appendiks E Modellering af sparse VAR(1) **63**
E.1 Modelvalidering af sparse VAR(1) for reaktor T14 . . . 65

Appendiks F Modellering af de enkelte OTU'er **74**

Appendiks G Hjælpesætninger **109**

1 Indledning

Biogas som en attraktiv energikilde

Fossile brændstoffer (kul, olie og naturgas) kan ved afbrænding anvendes som energi. I mange år har denne energikilde været anvendt, hvilket har haft store konsekvenser i form af globale klimaforandringer [2]. Afbrænding af de fossile brændstoffer har været en af de store syndere til den globale opvarmning, idet der ved afbrænding frigives store mængder af CO_2. Dette skaber en ubalance i kulstof kredsløbet, hvilket medfører, at meget af det frigivne CO_2 forbliver i atmosfæren. CO_2 i atmosfæren absorberer og udsender varmestrålinger, som skaber drivhuseffekten og dermed medvirker til den globale opvarmning [3]. Udover at der klimatisk er et problem ved anvendelse af fossile brændstoffer, er der også en risiko for, at vi løber tør for fossile brændstoffer [4]. I takt med at det globale energiforbrug stiger [5], er det derfor nødvendig at finde andre alternativer til at skaffe energi på.
En alternativ energikilde, som både er vedvarende og miljøvenlig, er biogas. Biogas består primært methan ($55\% - 80\%$) og kuldioxid ($20\% - 40\%$) og en mindre koncentration af nitrogen, hydrogen, svovlbrinte og oxygen [6]. Biogas dannes ved anaerob udrådning af gylle og andet organisk affald foretaget af mikroorganismer indenfor de to taksonomiske riger bakterier og arkæer [7]. Den anaerobe udrådning kan inddeles i fire processer, hvoraf de første tre processer foretages af bakterier imens den sidste proces foretages af arkæer[8]. Det tilbageværende bioslam (biproduktet af biogas-produktionen), indeholder en betydelig mængde af makro- og mikronæringsstoffer og anvendes derfor som gødning til landbruget [9]. Til sammenligning med husdyrgødning er fordelene ved anvendelse af bioslam, at det er fri for patogener, og at planterne har nemmere ved at optage nitrogen fra behandlet husdyrgødning[10]. Udover at biogas-produktionen er CO_2-neutral, bidrager produktionen også til en reduceret udledning af drivhusgasser såsom metan og dinitrogenoxid [11]. Alt dette gør, at produktionen af biogas er en attraktiv løsning til at skaffe energi på.
Biogasteknologien har i årtier været anvendt til at bortskaffe organisk affald, og er først i de seneste år blevet anvendt til at producere biogas med [12]. I Danmark blev regeringen i år 2009 enige om, at mindst 50% af gyllen skal udnyttes til biogas i år 2020 [13]. For blandt andet at imødekomme målet indgik regeringen i 2012 en aftale (energiaftalen [14]) om, at bevillige et tilskud til kraftværker eller virksomheder, der anvender biogas frem for naturgas [15]. Ifølge energistyrelsen rapport for vedvarende energi dækker biogas $15,28\%$ af vedvarende energi i DK i 2017 og er stadig i kraftig vækst [16].

Optimering af biogasproduktionen

Der er flere faktorer, der er med til at bestemme udbyttet af biogasproduktionen herunder substrat, temperatur, pH og partikelstørrelse. En ændring i en af faktorerne kan skabe en ubalance i det mikrobielle økosystem. Dette betyder, at mikroorganismerne først skal tilegne sig det nye miljø, hvilket derfor kan forsinke og formindske udbyttet af biogas. Tilegnelsen vil tage mindst 3 uger, hvis ændringen forekommer i enten temperatur eller substrat[7]. En tilstrækkelig variation i substratsammensætningen er

nødvendig for at imødekomme mikroorganismerne behov for næring. Manglende næring kan reducere/standse mikroorganismernes evne til reproduktion. Valget af substrat har afgørende betydning for mængden af produceret biogas og kvaliteten af bioslam[7][17]. Partikelstørrelsen på substratet har en betydning for hastigheden af den anaerobe udrådning, jo mindre partikler desto hurtigere sker processen, idet mindre partikler har et større overflade/volumen forhold, der gør, at mikroorganismerne nemmere kan angribe partiklerne [11].

Problemformulering

For at effektivisere biogasproduktionen skal faktorerne, der påvirker det mikrobielle samfund optimeres. Dog er viden omkring, hvordan faktorerne påvirker det mikrobielle samfund på nuværende tidspunkt minimal [8]. I mit speciale vil jeg undersøge den tidsafhængige mikrobielle dynamik, der er i et biogasanlæg ved at estimere en tidsserie over forekomsten af mikroorganismerne. Dette vil bringe ny viden indenfor biogasteknologien, da denne type analyse ikke er blevet lavet før. Modelleringen udføres ved hjælp af informationer omkring prøveserierne indsamlet fra et af Maarbjergs biogasanlæg [18], som behandler gylle.
Dataanalysen påbegyndes med rarefaction-kurver [19] for at se dækningsgraden [8]. Herefter udføres principal komponent analyse for at få et helhedsindtryk af, hvor forskellig stikprøverne er fra hinanden. Heatmap er anvendt til at se, hvor forskellig triplikaterne er fra hinanden. Spline er anvendt til at interpolere til ækvidistante tider, hvilket er nødvendig, når der skal estimeres en tidsserie. Sparse VAR(1) er anvendt for at få en ide om interaktionen mellem OTU'erne. En bedre model opnås ved at estimere en ny model, VARX(1), uden en straf (lasso-straf). Til sidst repræsenteres modellen, som en Lotka Volterra model.

1.1 Principal komponent analyse

I projektet arbejdes der med et datasæt bestående af mange variable (OTU'er). For at få et visuelt indtryk af, hvordan data opfører sig, ønskes der i projektet en lav dimensional præsentation af data, som opfanger meget af informationen gemt i data. Dette kan ofte lade sig gøre ved hjælp af principal komponent analyse. I dette afsnit præsenteres principal komponent analyse (PCA), som er et værktøj til at reducere et højt dimensionalt data, baseret på linear kombinationer, der har en høj varians. Afsnittet er baseret på kilderne [20] og [21]. Indledningsvis defineres principal komponent.

1.1 Definition (Principal komponent): Lad $\mathbf{x} = (x_1, x_2 \ldots, x_p)^T$ være en stokastisk multivariabel med $E[\mathbf{x}] = \boldsymbol{\mu}$ og $Cov[\mathbf{x}] = \Sigma$. Så kaldes $y_i = \mathbf{w}_i^T \mathbf{x}$ for den i'te principalle komponent, for $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{ip})^T$, hvis følgende betingelser er opfyldte for $i = 1, 2, \ldots, p$:

$$
\begin{aligned}
&||\mathbf{w}_i|| = 1 \\
&Cov(y_i, y_{j,j\neq i}) = 0 \\
&Var(y_i) \geq Var(y_{i-1})
\end{aligned}
$$

Elementerne i $\mathbf{w_i}$ kaldes loadings. ♦

Da principale komponenter er en linear kombination af variablerne i datasættet vil informationer ikke gå tabt, hvis analysen er baseret på alle principale komponenter. Af definition 1.1 fremgår det, at for et datasæt med p variable vil man opnå p principale komponenter. Disse komponenter er konstrueret således, at første komponent dækker den største varians i datasættet, anden komponent dækker den største varians af den tilbageværende varians osv. I tilfælde, hvor meget af variansen i et højt dimensionalt data er gemt i en lavere dimension vil få principale komponenter dække en stor del af variansen, hvorfor analysen kan baseres på de få komponenter. Ulempen ved at arbejde med principale komponenter fremfor variablerne fra datasættet er, at det kan være svært at drage konklusioner.
Der vil nu blive redegjort for estimaterne af loadings og varianserne for de principale komponenter. Idet kovariansmatricen er en symmetrisk matrix kan den skrives som $\Sigma = O^T D O$, hvor O er en ortogonal matrix med egenvektorer fra Σ og D er en diagonal matrix med egenværdierne fra Σ. Herefter ordnes søjlerne i O og egenværdierne således, at $\lambda_1 > \lambda_2 > \ldots, \lambda_p$. Så haves at:

$$
Var(y_i) = \mathbf{w}_i^T \Sigma \mathbf{w}_i = \mathbf{w}_i^T O^T D O \mathbf{w}_i = \tilde{\mathbf{w}}_i^T D \tilde{\mathbf{w}}_i = \tilde{w}_1^2 \lambda_1 + \tilde{w}_2^2 \lambda_2 + \cdots \tilde{w}_p^2 \lambda_p, \tag{1.1}
$$

hvor $\tilde{\mathbf{w}}_i = O\mathbf{w}_i$ er en rotation af $\mathbf{w}_i$. Af definition 1.1 dækker den første principale komponent den største varians. Derfor kan $\hat{\mathbf{w}}_1$ bestemmes ved at maksimere ligning (1.1) med hensyn til $\mathbf{w}_1$. Dette gøres ved først at maksimere højre siden af ligningen med hensyn til $\widetilde{\mathbf{w}}_1$ og derefter finde $\hat{\mathbf{w}}_1$:

$$
\widehat{\widetilde{\mathbf{w}}}_1 = \underset{\widetilde{\mathbf{w}}_1, ||\widetilde{\mathbf{w}}_1||=1}{\arg\max} \Big\{ Var(y_i) \Big\} = \mathbf{e}_i
$$

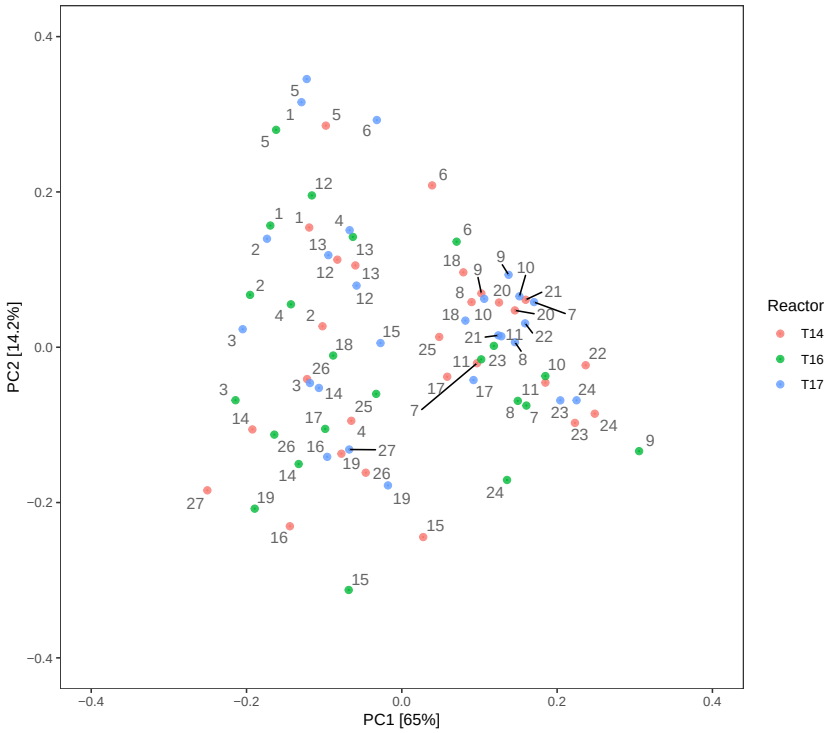
Da $\tilde{\mathbf{w}}_i = O\mathbf{w}_i$ er $\hat{\mathbf{w}}_1 = O^{-1}\widehat{\widetilde{\mathbf{w}}}_1 = O^T \mathbf{e}_1 = \mathbf{o}_1$, hvor $\mathbf{o}_1$ er den første række i den ortogonale matrix, som svarer til egenvektoren med størst egenværdi. Et estimat for $\mathbf{w_2}$ bestemmes på tilsvarende måde, men hvor man også har en begrænsning om, at $cov(y_1, y_2) = 0$.

$cov(y_1, y_2) = 0$ kan omskrives til:

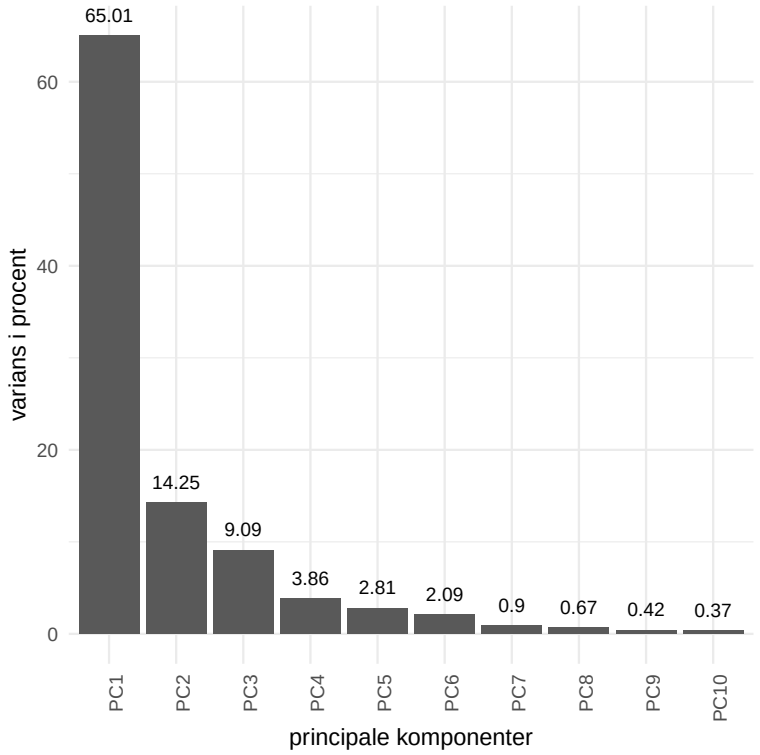
$$cov(y_1, y_2) = cov(\mathbf{w}_1\mathbf{x}, \mathbf{w}_2\mathbf{x}) = \mathbf{w}_1^T cov(\mathbf{x}, \mathbf{x})\mathbf{w}_2 = \tilde{w}_{11}\tilde{w}_{21}\lambda_1 + \tilde{w}_{21}\tilde{w}_{22}\lambda_2 + \cdots \tilde{w}_{1p}\tilde{w}_{2p}\lambda_p = 0, \tag{1.2}$$

Da $\widehat{\widetilde{\mathbf{w}}}_1 = \mathbf{e}_1$ resulterer ligning (1.2), at $\widetilde{w}_{21} = 0$. Dette betyder, at $\widehat{\widetilde{\mathbf{w}}}_1 = \mathbf{e}_2$ og dermed vil $\mathbf{o}_2$ dække den anden største varians. Analog kan det vises, at der for alle $i = 1, 2, \ldots, p$ gælder, at $\widehat{\widetilde{\mathbf{w}}}_i = \mathbf{o}_i$.
Afslutningsvis gives et eksempel, hvor der udføres principal komponent analyse baseret på et udsnit af datasættet. Formålet med eksemplet er at fremme forståelsen for principal komponent analyse og det bagvedliggende af et biplot med to principale komponenter, idet der under dataanalysen vil blive konstrueret et biplot.

1.2 Eksempel: Eksemplet tager udganspunkt i et udsnit af datasættet bestående af 14 udvalgte OTU'er observeret i de tre reaktorer. De 14 OTU'er er angivet i output 3.1. På figur 4 ses et pca biplot med første- og andet komponent. Ud fra biplottet er der en tendens til, at stikprøverne placerer sig i et mønster afhængig af tiden. Tilsammen forklarer første og andet komponent $79,26\%$ af den varians, der er i OTU'erne mellem stikprøverne. For at se, hvor stor en varians de øvrige komponenter udgør, er der lavet et screeplot, se figur 2. Hver bar på screeplottet viser, hvor stor en procentdel af den samlede varians komponenterne dækker. De første tre komponenter dækker $88,35\%$ af den samlede varians i datasættet. Man kan derfor udføre sin analyse på de første tre principale komponenter. Dette vil give en dimensionreduktion på $14 - 3 = 11$. Til at forstå, hvilke OTU'er, der hovedsageligt bidrager til variansen, ses der på loadings, se listing appendiks A.1. Dette viser, at variansen fra første komponent hovedsageligt udgøres af variationen fra OTU 1 og OTU 24. Variansen fra anden komponent udgøres af OTU 24 og OTU 13.



Figur 1: PCA biplot af stikprøverne taget i reaktor T14, T16 og T17 baseret på målinger for 14 udvalgte OTU'er.



Figur 2: Et screeplot, der viser, hvor stor en procentdel af den observerede varians hver principal kompenent beskriver.

1.2 Spline

Under dataanalysen anvendes en tredje grads spline til at interpolere abundansen af OTU'er til ækvidistatne tidspunkter. Nærværende afsnit vil give en forståelse for, hvad en spline er igennem dens definition og en udledning af et explicit udtryk for en d'te grads spline. Afslutningsvis analyseres, hvilke fordele der er ved at anvende splines gennem et eksempel. Afsnittet er baseret på kilderne [22],[23] og [24].

1.3 Definition (Spline): Lad $a < \zeta_1 < \zeta_2 < \cdots < \zeta_k < b$ og kald $\zeta_1, \zeta_2, \cdots, \zeta_k$ for knuder. En funktion $f_{d,k}(x)$ defineret på $[a,b]$ kaldes en d'te grads spline, hvis $f_{d,k}(x)$ er stykvise d'te grads polynomier i intervallerne $[a,\zeta_1], [\zeta_1,\zeta_2], \cdots, [\zeta_k,b]$, og opfylder at $f_{d,k}(x) \in C^{d-1}(a,b)$. ♦

At $f_{d,k}(x) \in C^{d-1}(a,b)$ betyder, at den $(d-1)$'te afledte eksisterer og er kontinuert i $[a,b]$. Med andre ord er funktionen kontinuert og funktionens afledte er kontinuerte op til orden $(d-1)$.

1.4 Sætning: En d'te grads spline $f_{d,k}(x)$ kan repræsenteres ved hjælp af basisfunktionerne $h_0(x), h_1(x), \cdots, h_{d+k}(x)$ som følgende:

$$f_{d,k}(x) = \sum_{l=0}^{d+k} \beta_l h_l(x), \tag{1.3}$$

hvor $h_j(x) = x^j$ for $j = 0, 1, \cdots, d$ og $h_{d+i} = (x-\zeta_i)_+^d = I[x \geq \zeta_i](x-\zeta_i)^d$ for $i = 1, 2, \ldots, k$ [23].

Bevis: Bevises fuldføres ved at vise, at ligning (1.3) opfylder betingelserne i definition (1.3). Det første betingelse er, at $f(x)$ er en stykvis d'te grads polynomium i hvert delinterval. Ved at skrive ligning (1.3) ud opnås:

$$f_{d,k}(x) = \begin{cases} f_{d,k,1}(x) = \beta_0 + \beta_1 x + \cdots \beta_d x^d, \text{ hvis } x \in [a,\zeta_1] \\ f_{d,k,2}(x) = \beta_0 + \beta_1 x + \cdots + \beta_d x^d + \beta_{d+1}(x-\zeta_1)_+^d, \text{ hvis } x \in [\zeta_1,\zeta_2] \\ f_{d,k,3}(x) = \beta_0 + \beta_1 x + \cdots + \beta_d x^d + \beta_{d+1}(x-\zeta_1)_+^d + \beta_{d+2}(x-\zeta_2)_+^d, \text{ hvis } x \in [\zeta_2,\zeta_3] \\ \vdots \\ f_{d,k,k+1}(x) = \beta_0 + \beta_1 x + \cdots + \beta_d x^d + \beta_{d+1}(x-\zeta_1)_+^d + \ldots + \beta_{d+k}(x-\zeta_k)_+^d, \text{ hvis } x \in [\zeta_k,b] \end{cases}$$

Af binomial expansion omskrives $(x-\zeta_i)^d = \sum_{m=0}^d \binom{d}{m} x^d \zeta_i^{d-m}$, hvorved det ses, at $f_{d,k}(x)$ er et stykvis d'te grads polynomium i hvert delinterval.
Hvis der for alle $c \in [a,b]$ gælder, at $lim_{x\to c} f_{d,k}(x) = f_{d,k}(c)$ og $lim_{x\to c} f_{d,k}^{(d-1)}(x) = f_{d,k}^{(d-1)}(c)$ er $f_{d,k}(x) \in C^{d-1}(a,b)$.
Først vises $lim_{x\to c} f_{d,k}(x) = f_{d,k}(c)$. Det er tydeligt, at ovenstående er gældende for alle $c \in x\backslash\{\zeta_1, \zeta_2, \cdots, \zeta_k\} \subseteq [a,b]$. For at vise, at funktionen også er kontinuert ved knuderne skal det gælde, at $f_{d,k,i+1}(\zeta_i) = f_{d,k,i}(\zeta_i)$. Lad $\zeta_i \in \{\zeta_1, \zeta_2, \cdots \zeta_k\}$, så opnås

$$\begin{aligned} f_{d,k,i+1}(\zeta_i) &= \sum_{m=0}^d \beta_0 \zeta_i^m + \beta_{d+1}(\zeta_i - \zeta_1)_+^d + \cdots + \beta_{d+i}(\zeta_i - \zeta_i)_+^d \\ &= \sum_{m=0}^d \beta_0 \zeta_i^m + \beta_{d+1}(\zeta_i - \zeta_1)_+^d + \cdots + \beta_{d+i-1}(\zeta_i - \zeta_{i-1})_+^d \\ &= f_{d,k,i}(\zeta_i). \end{aligned}$$

Differentieres funktionen op til den d'te afledte opnås følgende:

$$\begin{aligned}
f^{(1)}_{d,k}(x) =& \beta_1 + 2\beta_2 x + \cdots + d\beta_d x^{d-1} + d\beta_{d+1}(x-\zeta_1)^{d-1}_+ + \cdots + d\beta_{d+k}(x-\zeta_k)^{d-1}_+ \\
f^{(2)}_{d,k}(x) =& 2\beta_2 + 3\cdot 2\beta_3 x + \cdots + d(d-1)\beta_d x^{d-2} + d(d-1)\beta_{d+1}(x-\zeta_1)^{d-2}_+ \\
&+ \cdots + d(d-1)\beta_{d+k}(x-\zeta_k)^{d-2}_+ \\
f^{(d-1)}_{d,k}(x) =& (d-1)!\beta_{d-1} + d!\beta_d x + d!\beta_{d+1}(x-\zeta_1)_+ + \cdots + d!\beta_{d+k}(x-\zeta_k)_+ \\
f^{(d)}_{d,k}(x) =& d!\beta_d + d!(\beta_{d+1}I[x \geq \zeta_1] + \beta_{d+2}I[x \geq \zeta_2] + \cdots + \beta_{d+k}I[x \geq \zeta_k])
\end{aligned}$$

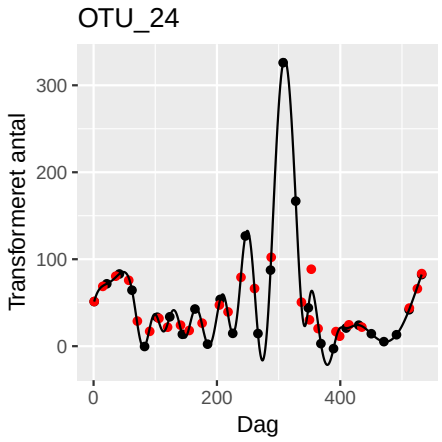
Når $x = \zeta_i$ er $(x-\zeta_i)_+ = 0$, hvilket betyder, at sidste led i $f^{(q)}_{d,k,i+1}(\zeta_i)$ går ud for $q = 1, 2, \cdots, d-1$, således at $f^{(q)}_{d,k,i+1}(\zeta_i) = f^{q}_{d,k,i}(\zeta_i)$. Nedenstående beregninger viser, at den $d'te$ afledte ikke er kontinuert, idet funktionsværdien springer med en afstand på $d!\beta_{d+i}$ ved knuderne.

$$\begin{aligned}
f^{d}_{d,k,i+1}(\zeta_i) &= d!\beta_d + d!(\beta_{d+1}I[x \geq \zeta_1] + \cdots + \beta_{d+i}I[x \geq \zeta_i]) \\
f^{d}_{d,k,i}(\zeta_i) &= d!\beta_d + d!(\beta_{d+1}I[x \geq \zeta_1] + \cdots + \beta_{d+i-1}I[x \geq \zeta_{i-1}])
\end{aligned}$$ ■

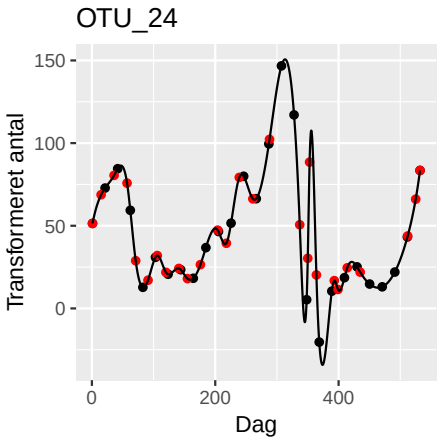
Ved at repræsentere en d'te grads spline som i ligning (1.3) kan $f_{d,k}(x)$ betragtes som den systematiske del af en lineær model, uden støj, med basisfunktionerne $h_l(x)$ som forklarende variable. Fordelen ved dette er, at sætninger samt propropositioner gældende for lineære modeller også gælder for en d'te grads spline, hvor designmatricen er $X = [(h_0(x_1), h_0(x_2), \cdots, h_0(x_n))^T, \cdots, (h_{d+k}(x_1), h_{d+k}(x_2), \cdots, h_{d+k}(x_n))^T] \in \mathbb{R}^{n\times(d+k+1)}$.

Når $n = d+k+1$ kan estimaterne for parametrene løses ved matrix-vektor ligningen $\mathbf{Y} = X\boldsymbol{\beta}$ imens der for $n > d+k+1$ haves, at $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}$ er entydigt bestemt, hvis $(X^TX)^{-1}$ eksisterer. Når $n < d+k+1$ findes der uendelig mange løsninger og et estimat for $\boldsymbol{\beta}$ kan ikke bestemmes entydigt. Med andre ord kan der maksimum vælges $n-4$ knuder for en tredje grads polynomium idet $n = 3+k+1$.
For at opnå en bedre forståelse for valg af grader og antal knuder for en spline gives et eksempel, hvor der fittes forskellige modeller til en delmængde af datasættet.

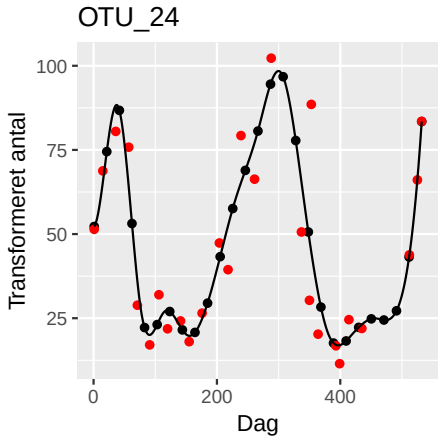
1.5 Eksempel: Eksemplet tager udgangspunkt i en delmængde af datasættet bestående af observationer for OTU 24 i reaktor $T14$ til ikke-ækvidistante tidspunkter. Ved hjælp af spline interpoleres estimater for abundansen af OTU'en til ækvidistante tidspunkter. Først undersøges hvilken grad, der fitter datasættet bedst. Til dette undersøges en 4. og en 3. grads spline som ses ved henholdsvis figur 3a og 3b. Af disse ses det, at en 4. grads spline peaker/topper mere end en 3. grads spline og følger derfor ikke samme tendens som datasættet, hvorfor det vurderes, at en kubisk spline passer bedst til datasættet. Herefter skal der findes en passende mængde knuder. Anvendes det maksimale antal knuder fluktuerer modellen for meget og dette vil være usandsynligt for abundansen af en mikroorganisme. Ydermere er der flere steder, hvor kurven bliver negativ. På figur 3c er der anvendt 8 knuder mindre. Disse knuder er valgt, så de er ligeligt fordelt mellem den mindste og største observation. For denne model ligger flere af de observerede udenfor modellen, men modellen ser mere stabil ud. Til sidst fittes en model med 15 knuder, men hvor knuderne af valgt ved hjælp af 15 percentiler. Dette vil gøre, at der vælges flere knuder, hvor der er observeret til flere tider. Dette giver god mening, idet man kan være mere sikker på, hvordan modellen ser ud, der hvor der er flere observationer, imens man er mere usikker på steder med få observationer.



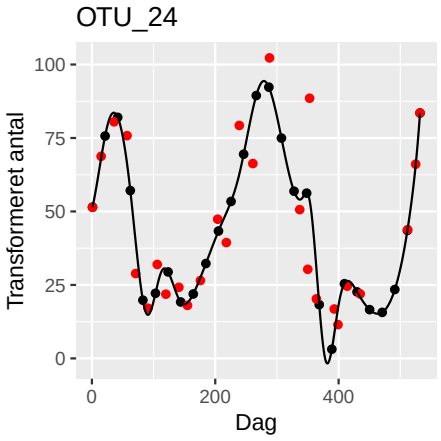
(a) Knuderne er valgt til at være de observerede tider fratrukket de tre mindste og tre største observerede tider.



(b) Knuderne er valgt til at være de observerede tider fratrukket de to mindste og to største observerede tider.



(c) Knuderne er valgt til at være 15 ækvidisstante tider fordelt mellem mindste og største tid.



(d) Knuderne er fundet ved hjælp af percentiler for $p \in \left\{\frac{1}{16}, \frac{2}{16}, \ldots, \frac{15}{16}\right\}$.

Figur 3: Kurverne viser spline for abundansen af OTU 24 i reaktor $T14$, som en funktion af antal dage fra første prøvetagningsdag. (a) er en 4. grads spline imens de øvrige er kubisk spline. De røde punkter markerer de observerede værdier, imens de sorte punkter markerer de estimerede værdier af abundansen til ækvidistante tidspunkter.

1.3 Lasso

Efter at have renset datasættet opnås et datasæt med 27 observationer og 14 variable, altså er antallet af observationer ikke meget større end antallet af variable. Konstrueres en model, hvor alle parametre indgår vil modelkompleksiteten være stor. Dette betyder, at modellen overfitter datasættet og dermed ikke er repræsentativt i forhold til et andet lignende datasæt. I dette kapitel introduceres en metode, kaldet lasso, til at straffe for modelkompleksiteten. Lasso er en velkendt *shrinkage method* for at regulere regressionskoefficienterne ned til nul. Dette vil resultere i en stigning af bias, men stigningen vil ofte være ubetydeligt i forhold til reduceringen af modelkompleksiteten. Afsnittet er baseret på kilderne [23], [25].

1.6 Definition (Lasso): Lad $i = 1, 2, \ldots, n$. For en lineær regression $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$, hvor ϵ_i er uafhængig identisk fordelt med $E[\epsilon_i] = 0$ defineres lasso estimaterne som

$$\hat{\boldsymbol{\beta}}_\lambda^L = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min}\Big\{ \sum_{i=1}^{n} \big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\big)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big\},$$

hvor $\lambda \sum_{j=1}^{p} |\beta_j|$ kaldes lasso straffen og λ tuning-parameteren. ♦

Lassostraffen kan ved hjælp af ℓ_1 normen skrives som $\lambda\|\beta\|_1$, og denne notation vil fremadrettet blive anvendt. Analogt kan lassoestimaterne findes som en løsning til følgende optimeringsproblem med bibetingelse:

$$\hat{\boldsymbol{\beta}}_\lambda^L = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min}\Big\{ \sum_{i=1}^{n} \big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\big)^2 \Big\} \quad \text{begrænset til} \quad \|\beta\|_1 \leq s.$$

Hvis s vælges til at være tilstrækkelig stor vil estimering af $\boldsymbol{\beta}^L$ ikke være påvirket af begrænsning, og derfor svarer $\hat{\boldsymbol{\beta}}^L$ til estimatet fundet ved hjælp af mindste kvadraters metode, $\boldsymbol{\beta}^{OLS}$. Tilsvarende gælder det, at hvis λ er 0 opnås $\boldsymbol{\beta}^{OLS}$. Vælges derimod $s = 0$ eller $\lambda \to \infty$ vil $\hat{\boldsymbol{\beta}}^L = \mathbf{0}$. For at få en bedre forståelse for lasso bestemmes et eksplicit udtryk for $\hat{\beta}_k^{Lasso}$.

1.7 Sætning: For en lineære regression, $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$, hvor $i = 1, 2, \ldots, n$, er lasso estimaterne for $k \in \{1, 2, \ldots, p\}$ givet ved:

$$\hat{\beta}_k^{Lasso} = \begin{cases} \dfrac{c_k - \lambda}{a_k} & \text{hvis} \quad c_k > \lambda \\ 0 & \text{hvis} \quad c_k \in [-\lambda; \lambda] \\ \dfrac{c_k + \lambda}{a_k} & \text{hvis} \quad c_k < -\lambda \end{cases} \tag{1.4}$$

hvor $c_k = 2\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1, j\neq k}^{p} \beta_j x_{ij}) x_{ik}$ og $a_k = 2\sum_{i=1}^{n} x_{ik}^2$.

Bevis: Beviset udføres ved at løse ligning (1.5). Ofte løses et optimumsproblem ved at finde den afledte. Dette dog ikke gøres i dette tilfælde, da den absolutte værdi ikke er differentiabel i 0. At den absolutte værdi ikke er differentiabel i nul ses ved følgende:

$$\lim_{\beta_j \to 0^+} \frac{|0 + \beta_j| - |0|}{\beta_j} = 1 \neq \lim_{\beta_j \to 0^-} \frac{|0 + \beta_j| - |0|}{\beta_j} = -1.$$

Optimumsproblemet løses i stedet ved brug af subdifferentialet(se definition G.1). Hvis 0 er i subdifferentialet, eksisterer der et minimum:

$$0 \in \frac{d}{d\beta_k}\big(RSS(\boldsymbol{\beta}) + \lambda\|x\|_1\big). \tag{1.5}$$

Først bestemmes $\frac{d}{d\beta_k}RSS(\boldsymbol{\beta})$:

$$\begin{aligned}\frac{d}{d\beta_k}RSS(\boldsymbol{\beta}) &= \frac{d}{d\beta_k}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1,j\neq k}^{p}\beta_j x_{ij} - \beta_k x_{ik})^2 \\ &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1,j\neq k}^{p}\beta_j x_{ij} - \beta_k x_{ik})x_{ik} \\ &= 2\sum_{i=1}^{n}\beta_k x_{ik}^2 - 2\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1,j\neq k}^{p}\beta_j x_{ij})x_{ik}\end{aligned}$$

Lad $c_j = 2\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1,j\neq k}^{p}\beta_j x_{ij})x_{ik}$ og $a_k = 2\sum_{i=1}^{n}x_{ik}^2$. Subdifferentialet, til lasso straffen bestemmes til:

$$\frac{d}{d\beta_k}\lambda\|\boldsymbol{\beta}\|_1 = \begin{cases}\lambda & \text{hvis} \quad \beta_k > 0 \\ [-\lambda,\lambda] & \text{hvis} \quad \beta_k = 0 \\ -\lambda & \text{hvis} \quad \beta_k < 0\end{cases}$$

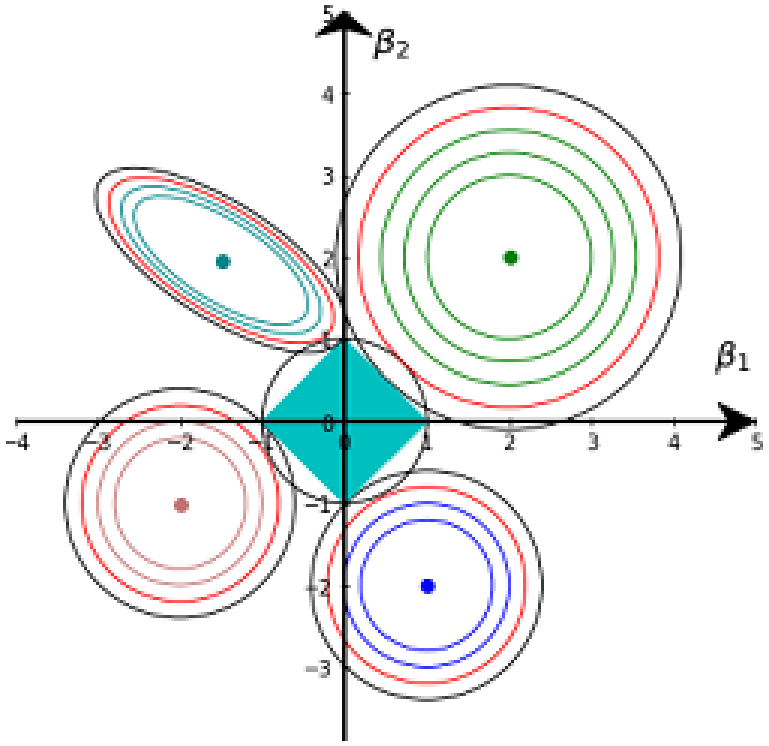
Heraf opnås:

$$\frac{d}{d\beta_k}(RSS(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1) = \begin{cases}-c_k + \beta_k a_k + \lambda & \text{hvis} \quad \beta_k > 0 \\ \{-c_k - \lambda; -c_k + \lambda\} & \text{hvis} \quad \beta_k = 0 \\ -c_k + \beta_k a_k - \lambda & \text{hvis} \quad \beta_k < 0\end{cases} \tag{1.6}$$

Ved at sammenholde (1.6) og (1.5) opnås (1.4). ■

Afslutningsvis gives et eksempel på mulige løsninger til lasso-estimatet for en lineær regression med to forklarende variable. Formålet med eksemplet er at vise illustrativt, at lasso i de fleste tilfælde vil regulere nogle af regressionskoefficienterne ned til nul. For at fremme forståelsen af lasso straffen sammenlignes i eksemplet også med en anden straf kaldet ridgeregression. Den eneste forskel der er mellem ridgeregression og lasso-straffen er, at begrænsning for ridgestraffen svarer til $\sum_{j=1}^{p}\beta_j^2 \leq s$.

1.8 Eksempel: På figur 4 er niveaukurverne for objektfunktionen svarende til $RRS(\beta_1, \beta_2)$ indtegnet for fire forskellige datasæts, hvilket ses som ellipser og cirkler på figuren. Begrænsningsregionen af lasso-straffen er det blå-farvede kvadrat i midten med $s = 1$. Cirklen i midten med centrum i nul svarer til begrænsningsregionen for ridgeregression med $s = 1$. Optimumspunkterne findes ved skæringspunkterne mellem niveaukurven og begrænsningsregionen. Ud fra figuren ses det, at niveau-kurverne er lettere tilbøjelig til at ramme et af hjørnepunkterne af kvadratet, hvor en af parametrene er nul, frem for en af dens sider. Til sammenligning med begrænsningsregionen for ridgeregression er det sjældent, at niveaukurverne første gang rammer cirklen ved en af akserne. Lasso-straffen er derfor anvendelig, når en delmængde af variablene skal vælges.



Figur 4: På figuren er indtegnet niveaukurver for $RRS(\boldsymbol{\beta})$ for fire forskellige datasæt. Kvadratetet i midten er begrænsningsmængden for lasso. Cirklen med centrum i nul er begrænsningsmængden for ridgeregression. Billedet er hentet fra [26].

1.4 Tidsrækkeanalyse

I projektet arbejdes der med et datasæt bestående af observerede OTU'er i et biogasanlæg henover tiden. Da et mikrobielt samfund er et netværk af mikroorganismer vil en fjernelse/reducering af en mikroorganisme kunne have betydning for de øvrige mikroorganismer. Ved hjælp af en multivariat tidsserie, kaldet vektorautoregressiv tidsserie, undersøges om dette er gældende for OTU'er i et biogasanlæg. I nuværende afsnit præsenteres vektor autoregressive tidsserier med og uden eksogene variable, da disse modeller anvendes i dataanalysen. Indledningsvis defineres en tidsserie. Kilderne anvendt til afsnittet er [27] og [28].

1.9 Definition: En stokastisk proces $\{x_t; t \in T\}$ kaldes en tidserie, hvor indeksmængden T repræsenterer tiden t, som er diskret og ækvidistant. ⧫

Observeres en hændelse $x \in \mathbb{R}$ over en periode $t \in T$ haves en realisering af tidsserien $\{x_t; t \in T\}$.

1.10 Definition: En multivariet stokastisk proces $\{\mathbf{x}_t; t \in T\}$ kaldes en multivariat tidsserie, hvis indeksmængden T repræsenterer tiden t. ⧫

Komponenterne i en multivariat tidsserie svarer således til $\{x_{ti}\}$ for $i = 1, 2, \ldots m$. Ofte når der arbejdes med tidsserier er det kun en realisering af tidsserien, som er tilgængelig. Med kun en realisering er det svært at beskrive en tidsserie, dog kan dette lade sig gøre ved at antage, at der eksisterer en form for regelmæssighed i fordelingen af tidsserien. I det følgende defineres stationaritet, som er et begreb for eksistensen af regelmæssighed i tidsserien.

1.11 Definition (Strengt stationær): En tidsserie $\{x_t; t \in T\}$ er strengt stationær, hvis den simultane fordeling er uændret under translation af tiden. ⧫

1.12 Definition (Svag stationaritet): En tidsserie $\{x_t; t \in T\}$ er svag stationær, hvis følgende betingelser er opfyldte for alle $t \in T$:

$$E[x_t] = \mu$$
$$\text{Var}[x_t] = \sigma^2 < \infty$$
$$\text{Cov}(x_{t_i}, x_{t_{j,i\neq j}}) \text{ afhænger kun af } h = |t_i - t_j|.$$

I stedet for at antage strengt stationaritet antages ofte svagt stationaritet, da det er svært statistisk at påvise strengt stationaritet. Antages strengt stationaritet om abundansen af OTU'er i biogasanlægget vil det betyde, at sandsynligheden for at observere en given OTU er den samme hele året, hvilket er særdeles usandsynligt for alle OTU'er, da gyllens artssammensætning ændres i løbet af året. I praksis er det derfor sjældent, at fordelingen af en tidsserie er uændret over tiden.
En almindelig anvendt multivariat tidsserie er en vektor autoregressiv model, VAR(p).

1.13 Definition: En multivariate tidsserie x_t følger en vektorautoregressiv model af orden p, VAR(p), hvis

$$\mathbf{x}_t = \boldsymbol{\phi}_0 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \mathbf{x}_{t-i} + \mathbf{w}_t,$$

hvor $\boldsymbol{\phi}_0$ er en k-dimensional konstant vektor, $\boldsymbol{\Phi}_i$ er $k \times k$ matricer for $i > 0$, $\boldsymbol{\Phi}_p \neq \mathbf{0}$ og $\mathbf{w}_t$ er normalfordelte vektorer der er uafhængige af hinanden. ♦

Ækvivalent kan VAR(p) ved brug af lag operatoren formuleres som $\boldsymbol{\Phi}(B)(\mathbf{x}_t - \boldsymbol{\mu}) = \mathbf{w}_t$, hvor $\boldsymbol{\Phi}(B) = I - \boldsymbol{\Phi}_1 B - \boldsymbol{\Phi}_2 B^2 - \ldots - \boldsymbol{\Phi}_p B^p$, hvor der gælder, at VAR(p) er stationær, hvis rødderne for den karakteristiske ligning $\det(\Phi(B)) = 0$ ligger udenfor enhedscirklen. For en $VAR(1)$ er dette ækvivalent med at den absolutte værdi af egenværdierne for $\boldsymbol{\Phi}$ er mindre end en. Dette udledes ved først, at multiplicere den karakteristiske ligning, $\det(I - \boldsymbol{\Phi}B) = 0$, med B^{-k} på begge sider, som resulterer i den karakteristiske ligning af $\boldsymbol{\Phi}$, $\det(\frac{1}{B}I - \boldsymbol{\Phi}) = 0$. Da egenværdierne for Φ er givet ved alle værdier af $\frac{1}{B}$, der opfylder den karakteristiske ligning, er stationaritet for VAR(1) opfyldt, hvis den absolutte værdi af egenværdierne er mindre end en.
Bemærk at VAR(p) kan opstilles som en multipel linear regression, hvor de forklarende variable er responsvariablen til p tider tilbage, $\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots, \mathbf{x}_{t-p}$. Dette betyder, at sætninger og propositioner gældende for en multipel linear regression også gælder for en VAR(p). For en 14-dimensional tidsserie er en VAR(1) givet som

$$\begin{bmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{14t} \end{bmatrix} = \begin{bmatrix} \phi_{1,0} \\ \phi_{2,0} \\ \vdots \\ \phi_{14,0} \end{bmatrix} + \begin{bmatrix} \phi_{1,11} & \phi_{1,12} & \phi_{1,13} & \cdots & \phi_{1,114} \\ \phi_{1,21} & \phi_{1,22} & \phi_{1,23} & \cdots & \phi_{1,214} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{1,141} & \phi_{1,142} & \phi_{1,143} & \cdots & \phi_{1,1414} \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \\ \vdots \\ x_{14,t-1} \end{bmatrix} + \begin{bmatrix} w_{1t} \\ w_{2t} \\ \vdots \\ w_{14t} \end{bmatrix},$$

hvor $\boldsymbol{\phi}_0 = (I - \boldsymbol{\phi}_1)\boldsymbol{\mu}$. VAR(1) modellen er den simpleste multivariate tidsserie. Har man ingen ide om, hvilken model der fitter ens data, kan man derfor starte med en VAR(1) for at se, hvilken struktur der er. Ofte opdages, at nogle af de univariate tidsserier kun afhænger af sig selv, men som samtidig har indflydelse på en af de øvrige tidsserier. I dette tilfælde kan man fjerne den pågældende tidsserie og betragte tidsserien som en eksogen variabel. Modellen kaldes da for en vektor autoregressiv model med eksogene variable.

1.14 Definition: En multivariat tidsserie, $\mathbf{x}_t$, følger en vektor autoregressiv model med eksogone variable, VARX(p), hvis

$$\mathbf{x}_t = \phi_0 + \Gamma \mathbf{u}_t + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \mathbf{x}_{t-j} + \mathbf{w}_t,$$

hvor ϕ_0 er en $k \times 1$ vektor, Γ er en $k \times r$ matrice, $\mathbf{u}_t$ en $r \times 1$ vektor af eksogene variable, $\boldsymbol{\Phi}_j$ er $k \times k$ overgangsmatricer og $\mathbf{w}_t$ er normalfordelte, uafhængige vektorer. ♦

Ved en sammenligning af definitionerne for henholdsvis VAR(p) og ARX(p) fremgår det, at VAR(p) er et specialtilfælde er VARX(p).

1.5 Lotka volterra model

Lotka volterra modellen blev introduceret omkring år 1950, og siden hen har man anvendt denne model i utallige sammenhænge. Oprindeligt blev det introduceret for at beskrive interaktionen mellem rovdyr og byttedyr. Siden hen har man udvidet modellen på forskellig vis, afhængig af, hvad der beskrives [29]. Inspireret af modellen beskrives dynamikken mellem OTU'erne ved hjælp af den udvidede Lotka-volterra model kaldet generaliseret Lotka-volterra model. I projektet bestemmes estimaterne for parametrene i Lotka-Volterra modellen ved hjælp af estimaterne bestemt for VARX(p). I nærværende afsnit redegøres for et estimat af parametrene for Lotka-Volterra modellen ud fra estimaterne fra VARX(p).

1.15 Definition (Generaliseret Lotka Volterra): For $i = 1, 2, \ldots, L$ er den generaliserede Lotka Volterra model givet ved L uafhængige førsteordens differentialligninger:

$$\frac{d}{dt}x_i(t) = \alpha_i x_i(t) + x_i(t)\sum_{j=1}^{L} M_{ij}x_j(t), \tag{1.7}$$

hvor $x_i(t)$ angiver koncentrationen af den i'te art til tidspunktet t, α_i angiver den i'te arts væksthastighed og M_{ij} angiver størrelsen på interaktionen af art j på art i [30]. ♦

Bemærk, at i definitionen kræves, at t er kontinuert, idet man ellers ikke kan udregne den afledte. Dog er observationerne i datasættet observeret til diskrete tidspunkter. Der ses derfor på, hvordan den generaliserede Lotka Volterra model kan omskrives, så den gælder for diskrete tider. Først divideres ligning (1.7) med $x_i(t)$, hvorved følgende udtryk opnås:

$$\frac{d}{dt}\ln(x_i(t)) = \alpha_i + \sum_{j=1}^{L} M_{ij}x_j(t), \tag{1.8}$$

når $x_i(t) \neq 0$. Lad $t_k \in \{t_1, t_2, \ldots, t_n\}$ være diskrete tidspunkter, så kan venstre siden af ligning (1.8) skrives som:

$$\begin{aligned}\frac{d}{dt}\ln(x_i(t))\bigg|_{t=t_k} &\approx \frac{\ln(x_i(t_{k+1})) - \ln(x_i(t_k))}{t_{k+1} - t_k} = \ln(x_i(t_{k+1})) - \ln(x_i(t_k)) \\ &\approx \ln(\mu_i) + \frac{1}{\mu_i}(x_i(t_{k+1}) - \mu_i) - \Big(\ln(\mu_i) + \frac{1}{\mu_i}(x_i(t_k) - \mu_i)\Big) \\ &= \frac{1}{\mu_i}(x_i(t_{k+1}) - x_i(t_k)),\end{aligned}$$

Hvor første lighed følger af, at tidspunkterne er taget til ækvidistante tider, hvor $t_{k+1} - t_k$ vælges som en tidsenhed. Anden approksimation følger af tangentens ligning. Sammenholdes dette resultat med (1.8) haves:

$$\begin{aligned}\frac{1}{\mu_i}(x_i(t_{k+1}) - x_i(t_k)) =& \alpha_i + \sum_{j=1}^{L} M_{ij}x_j(t_k) \\ \Leftrightarrow x_i(t_{k+1}) =& x_i(t_k) + \mu_i + \mu_i\alpha_i \sum_{j=1}^{L} M_{ij}x_j(t_k) \\ =& x_i(t_k) + \mu_i\alpha_i + \mu_i\mathbf{m}_i^T\mathbf{x}(t_k)\end{aligned} \tag{1.9}$$

Et udtryk for α_i og $\mathbf{m}_i$ findes ved at sætte ligning (1.9) lig med VAR(1):

$$x_i(t_k) + \mu_i\alpha_i + \mu_i\mathbf{m}_i^T\mathbf{x}(t_k) = \boldsymbol{\phi}_{0_i} + \boldsymbol{\Phi}_{1_i}\mathbf{x}(t_k),$$

Dette resulterer i at $\alpha_i = \dfrac{\boldsymbol{\phi}_{0_i}}{\mu_i}$ og $\mathbf{m}_i = (\boldsymbol{\Phi}_{1_i} - \mathbf{e}_i)\dfrac{1}{\mu_i}$.

1.6 Modelvalidering

Krydsvalidering

Dette afsnit er baseret på [23]. Krydsvalidering er en *resampling method*, som i projektet er anvendt til bestemmelse af størrelsen for lassostraffen (λ) for en sparse VAR(1). Krydsvalidering foregår ved at dele observationerne tilfældigt op i k (approksimativ) lige store mængder. Så bestemmes estimaterne for en given model ud fra $k-1$ mængder, kaldet træningssæt. Når estimaterne er bestemt undersøges, hvor langt væk observationerne fra den tilbageværende mængde, kaldet valideringssæt, ligger fra den fittede model. Et mål for hvor langt væk observationerne ligger fra den fittede model er den middel kvadrerede fejl (MSE) som beregnes i krydsvalidering. Herefter bestemmes estimaterne igen, men med et andet træningssæt og valideringssæt. Denne proces gentages i alt k gange således, at hver af de inddelte mængder er blevet anvendt som valideringssæt. For hver gang beregnes middel kvadrerede fejl, baseret på valideringssæt, som anvendes til at bestemme krydsvalidering estimatet:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

Ved at estimere en sparse VAR(1) med forskellige grader af lassostraffen (λ) kan man for hver estimering af VAR(1) beregne krydsvalideringsestimatet for så at vælge modellen, der har det laveste krydsvalideringsestimatet. Der ses nu på betydningen af k. Det største værdi k kan antage er n. I dette tilfælde kaldes krydsvalideringen for *leave-one-out cross-validation*. Når $k = n$ vil de fittede modeller minde meget om hinanden, idet modellerne er estimeret på basis af samme observationssæt med undtagelse af en observation. Dette vil betyde, at MSE_i vil være stærkt korrelerede og dermed vil middelværdien af MSE_i ikke være repræsentativ i forhold til et andet datasæt. Modsat vil en lille værdi af k lettere resultere i overfit af middelværdien for MSE_i, idet træningssættet er baseret på en lille delmængde af datasættet. I mange anvendelser vælges k ofte til 5 eller 10. Når der i praksis skal vælges et k er det vigtigt, at man tænker over, hvor stort ens datasæt er, idet valget af k kan have større betydning for et lille datasæt.

ACF og PACF korrelogrammer

Dette afsnit er baseret på kilderne [28] og [31]. ACF- og PACF korrelogrammer anvendes i projektet til modelvalidering. ACF og PACF korrelogrammer er en visuel præsentation af henholdsvis autokorrelation funktion (ACF) og partiel autokorrelation funktion (PACF) op til et bestemt lag. For at forstå, hvordan disse er anvendt, er det nødvendigt at forstå ACF og PACF. Derfor præsenteres først en definition for korrelationerne, hvorefter der argumenteres for, hvordan man kan bruge korrelogrammerne til modelvalidering.

1.16 Definition (Autokovariansen og autokorrelation): Lad x_t være en tidsserie og $t_i, t_j \in t$. Så er autokovariansen funktionen og autokorrelationen funktionen defineret ved henholdsvis

$$\gamma_x(i,j) = cov(x_{t_i}, x_{t_j}) = E[(x_{t_i} - E[x_{t_i}])(x_{t_j} - E[x_{t_j}])]$$
$$\rho_x(i,j) = corr(x_{t_i}, x_{t_j}) = \frac{\gamma_x(i,j)}{\sqrt{\gamma_x(i,i)\gamma_x(j,j)}}$$

For en svag stationær proces gælder det, at kovariansen kun afhænger tidsafstanden, $h = |t_i - t_j|$, og derfor kan kovariansen noteres som $\gamma(h)$. Ved hjælp af denne notation

og viden om, at variansen er konstant for en tidsserie, er korrelationen for en stationær proces givet ved $\rho_x(h) = \dfrac{\gamma(h)}{\gamma(0)}$.

1.17 Definition (Partielle autokorrelations funktion): Den partielle autokorrelation funktion (PACF) for en stationær tidsserie x_t for $h = 1, 2 \ldots$ er givet ved

$$\phi_{11} = corr(x_{t+1}, x_t) = \rho(1)$$
$$\phi_{hh} = corr(x_{t+h} - \hat{x}_{t+h}, x_t),$$

hvor $\hat{x}_{t+h}$ er den lineære prædiktor af x_{t+h} baseret på $x_{t+1}, \ldots, x_{t+h-1}$. ♦

PACF til lag h skal derfor forstås som korrelation mellem x_t og x_{t-h}, hvor man har fjernet effekten af korrelationen fra $x_{t+1}, \ldots, x_{t+h-1}$. Disse kan anvendes både i start- og slutprocessen af modelleringen. I startprocessen kan de anvendes til at opnå en ide om hvilken model, der vil fitte datasættet godt. Imens man i slutprocessen kan bruge ACF og PACF korrelogrammerne for residualerne til at se, om de medtagede lags beskriver association mellem prædiktoren og de tidligere lags tilstrækkeligt. Afslutningsvis gives et eksempel, hvor ACF og PACF korrellogrammer anvendes til bestemmelse af lags for en ARX.

1.18 Eksempel: Eksemplet tager udgangspunkt i det rensede datasæt bestående af abundanserne for de 14 OTU'er. Der konstrueres først en ARX(1) for abundansen af OTU 7 med abundansen af OTU 1 som eksogen variabel.

```
model_otu7=arima(T14_data_matrix[,11], c(1,0,0), xreg=T14_data_matrix[c(NA,1:26),7])
par(mfrow=c(1,2))
model_otu7;ptest(model_otu7$coef/sqrt(diag(model_otu7$var.coef)))[c(1,3)]
```

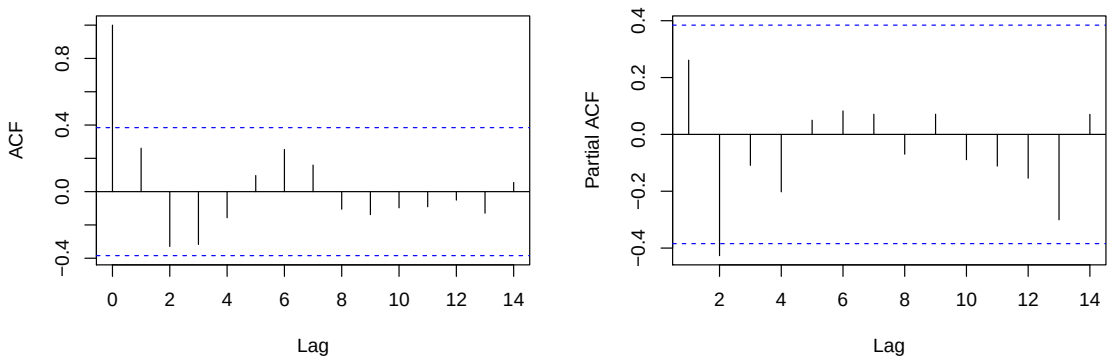
```
Call:
arima(x = T14_data_matrix[, 11], order = c(1, 0, 0), xreg = T14_data_matrix[c(NA,
    1:26), 7])

Coefficients:
         ar1  intercept  T14_data_matrix[c(NA, 1:26), 7]
      0.5942    58.0817                          -0.0969
s.e.  0.1923     6.4912                           0.0415

sigma^2 estimated as 16.22:  log likelihood = -73.33,  aic = 154.66
                          ar1 T14_data_matrix[c(NA, 1:26), 7]
                  0.001995528                     0.019716287
```

På figur 5 er ACF og PACF korrelogrammerne vist. Det observeres, at PACF for lag 2 er signifikant, hvilket kan indikere at en ARX(2) vil fitte bedre til datasættet. Der fittes derfor en ARX(2).

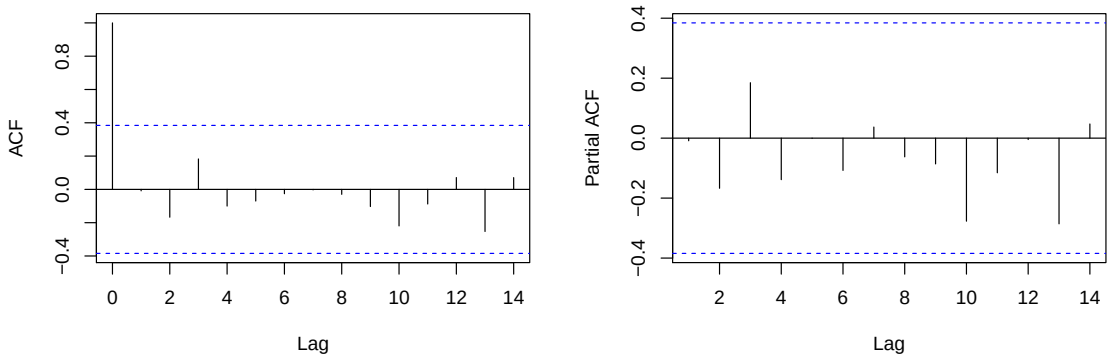
```
model_otu7.2=arima(T14_data_matrix[,11], c(2,0,0), xreg=T14_data_matrix[c(NA,1:26),7])
model_otu7.2;ptest(model_otu7.2$coef/sqrt(diag(model_otu7.2$var.coef)))[c(1,2,4)]
```

```
Call:
arima(x = T14_data_matrix[, 11], order = c(2, 0, 0), xreg = T14_data_matrix[c(NA,
    1:26), 7])

Coefficients:
         ar1      ar2  intercept  T14_data_matrix[c(NA, 1:26), 7]
      0.8720  -0.6652    53.7175                          -0.0683
s.e.  0.1673   0.1596     4.2545                           0.0275

sigma^2 estimated as 10.09:  log likelihood = -67.69,  aic = 145.38
                          ar1                       ar2
                  1.877606e-07              3.064952e-05
T14_data_matrix[c(NA, 1:26), 7]
                  1.319614e-02
```

På figur 5 er korrelogrammerne af residualerne for en ARX(2) for OTU 7 vist. Sammenlignes figur 5 med figur 6 ses det, at den signifikante korrelation for lag 2 forsvinder, når der fittes en ARX(2). I dette tilfælde vil ARX(2) derfor være en bedre model end ARX(1).



Figur 5: ACF og PACF af residualerne for ARX(1) for OTU 7.



Figur 6: ACF og PACF af residualerne for ARX(2) for OTU 7.

2 Metode

Det indeværende kapitel rummer en redegørelse for de metoder, som er anvendt for at besvare problemformuleringen. Da metodevalg afhænger af hvilket data, der arbejdes med, indledes kapitlet med en beskrivelse af data.

2.1 Databeskrivelse

I projektet arbejdes der med data for bakterier fra et af Maabjerg's biogasanlæg, som behandler gylle. Data er inddelt i to datasæt, kaldet `otu_bak` og `meta`. `otu_bak`, som indeholder informationer om, hvilke OTU'er der er observeret i stikprøverne, antallet af sekvenser observeret for hver OTU og en klassifikationen af sekvenserne.
Prøverne er taget fra tre forskellige reaktorer ($7500m^3$ cylinder med omdrejninger) i et biogasanlæg, kaldet T14, T16 og T17, hvoraf T14 og T16 er placeret parallelt overfor hinanden imens T17 er placeret modsat. Indsamlingen er foretaget på 27 dage ujævnt fordelt over en periode af cirka 1,5 år (25.06.2015-07.12.2016). De første 14 prøvetagninger blev foretaget ved at udføre triplikater, hvorefter dette er blevet stoppet grundet økonomiske årsager. Til sekventering af DNA-sekvenser er der udført gensekventering af 16S rRNA ved hjælp af next generation sekventering [32] for bakterier. Ved hjælp af algoritmen USEARCH7 er DNA-sekvenser med mindst 97% sekvensenshed klynget sammen til klynger, kaldet OTU'er, hvorefter en sekvens vælges som repræsentativ for hver OTU. OTU'erne er klassificeret ved hjælp databasen silva rRNA database project med den nyeste udgivelse, silva 132, den 13-12-2017 [33]. Det rå datasæt `OTU_bakterier` består af 3375 OTU'er, fordelt over 174 stikprøver, med tilhørende klassifikationer af OTU'er på de syv taksonomiske niveauer fra rige til art. I det rå datasæt `Meta_bakterier` indgår 5 forklarende variable, se Tabel 1.

Tabel 1 Variable i metadata

Reactor Angiver hvilken reaktor prøven er taget fra. Kan antage værdierne $\{T14, T16, T17\}$	Collection_Date Angiver dato for prøvetagning.	Data_Point Angiver prøvetagningsdag og kan antage værdierne $\{1, \cdots, 27\}$	Week Angiver uge for prøvetagning	Day Angiver en nummering af dage fra første prøvetagning som nr. 1. Kan antage værdier indenfor intervallet $[1, 532]$

Uoverensstemmelser i data

I dette afsnit tolkes der samlet set på informationerne i datasættene ved blandt andet at sammenligne informationer om en observation på tværs af variablene.
Under variablen `Datapoint` i `meta` står der angivet, at to af stikprøverne, 16SAMP-9801

og 16SAMP-9840, er indsamlet på prøvetagningsdag 1, hvilket er i modstrid med de øvrige variable, som overbevisende indikerer, at prøverne er indsamlet på prøvetagningsdag 2. Det antages derfor, at der er sket en tastefejl, og `Datapoint` ændres til 2 for de to stikprøver. I `meta` indgår tre prøver kaldet kontrolprøver, altså prøver der ikke er udtaget i biogasanlægget, men indeholder kendte koncentrationer, der er anvendt som en del af sekventeringsprocessen, hvorfor kontrolprøverne med tilhørende OTU'er fjernes fra datasættene. Herved reduceres antallet af OTU'er til 3344.
Yderligere indgår der to triplikater, prøverne er taget på prøvetagningsdag 7 og 12 i henholdsvis $T14$ og $T17$, som man har sekventeret over to omgange, da resultatet ikke var godt under første sekventering, se appendiks A for et udsnit af data. Hvilken af de to triplikater, der er bedst og dermed beholdes, argumenteres der for i næste afsnit. I amplicon sekventering indgår en masse trin, hvor der nemt kan opstå fejl[34]. Haves eksempelvis en stikprøve med få reads, er det ofte en indikation af, at der har været noget galt ved behandlingen af stikprøven [34]. I dette tilfælde må man sekventere en gang til eller udelade prøven til den videre analyse.
Det kommende afsnit beskriver de metoder, der er anvendt til at undersøge validiteten af de opnåede sekvenser.

2.2 Validering af sekvenserne

Til at undersøge, hvilke prøver der er valide og dermed skal beholdes, undersøges kvaliteten af prøverne ved hjælp af rarefaction kurver, variationen af reads pr. prøve indenfor en triplikat og principal component analysis (PCA). Rarefaction kurver viser, om de opnåede sekvenser dækker den observerede mikrobielle diversitet, hvorved der vælges et minimum grænse for antallet af sekvenser en prøve skal have. Ydermere undersøges variationen af antal reads indenfor hver triplikat. Er der en høj variation i antallet af reads, er det en indikation af, at kvaliteten af prøverne indenfor en triplikat varierer meget. PCA og heatmap anvendes til at se, om prøverne indenfor en triplikat ligner hinanden. Fordelen ved at bruge heatmap er, at man hurtigt kan sammenligne abundansen af OTU'er på tværs af prøverne. Er der en prøve inden for et triplikat, som ikke ligner de øvrige prøver, fjernes prøven.

Håndtering af triplikater

For hver triplikat vælges en prøve, som værende repræsentativ for observationen taget til tiden t. Udvælgelsen er baseret på at opnå prøven med størst kvalitet. Proceduren for udvælgelsen er følgende: Hvis alle prøver indenfor samme triplikat ligner hinanden vælges prøven med flest antal reads. Hvis to af prøverne indenfor samme triplikat ligner hinanden, vælges den prøve, ud fra de to prøver, som har flest antal reads. Er der ingen af prøverne, som ligner hinanden, fjernes triplikatet. Dette resulterer i at antallet af OTU'er reduceres til 3288 OTU'er.

2.3 Udvælgelse af OTU'er til modellering

Igennem projektet ønskes det at konstruere en model for de hyppigst forekommende OTU'er. Der undersøges derfor, hvilke OTU'er der er hyppigst forekommende i de enkelte reaktorer, hvorefter OTU'erne på tværs af reaktorerne sammenlignes. Hvilken metode, der er den optimale til at udvælge de hyppig forekommende OTU'er vides ikke og derfor afprøves forskellige metoder. Metoden der giver et tilstrækkeligt antal OTU'er, og som samlet har reads nok til at dække størstedelen af total antal reads i datasættet, vælges.
Før antallet af OTU'er reduceres, transformeres antallet af reads ved at tage kvadratroden.

Yderligere beregnes de relative frekvenser pr. prøve, hvilket gør, at antallet af reads er sammenlignelige på tværs af prøverne. Antallet af OTU'er, der har et maksimum, median og gennemsnit større end eller lig med $0,1\%$, $0,5\%$, 1% ses i henholdsvis tabel 2, 3 og 4. Fællesmængden samt foreningsmængden af OTU'erne for reaktorerne ses ved de samme tabeller. Fællesmængden er beregnet for at finde de OTU'er, der er repræsentative for hele biogasanlægget. Procenterne i parentesen angiver, hvor stor en procentdel antallet af reads fra delmængden af OTU'er udgør fra summen af reads i datasættet. R-koderne anvendt til at beregne værdierne i tabellerne ses i A.3. Modellen, der skal konstrueres, afhænger af tiden. Stikprøverne er observeret i $23-27$ tidspunkter, se kapitel 3. Konstrueres en model med $23-27$ parametre for hver reaktor vil modellerne være overfittet, hvorfor ønsket er, at reducerer antallet af OTU'er til mindre end 23. Modsat kan modellen ikke anvendes til at sige noget om de mikrobiologiske forhold, hvis modellen kun er baseret på få OTU'er. Metoderne der resulterer i mindre end 23 OTU'er og forekommer i alle reaktorer, er når max $= 1\%$, median $\geq 0,5\%$, median $\geq 5\%$, gennemsnit $\geq 1\%$ og gennemsnit $\geq 0,5\%$. To af de nævnte metoder resulterer i 5 OTU'er, hvilket vurderes at være for få bl.a. på grund af deres dækningsgrad. Gennemsnittet $\geq 0,5\%$ har en dækningsgrad på cirka 5% mere end median $\geq 0,5\%$ og maksimum $\geq 1\%$, men består af op til 5 OTU'er mere, hvorfor denne metode udelukkes. Max$= 1\%$ og median $\geq 0,5\%$ dækker næsten den samme procentdel, men da medianen er et bedre estimat end maximum vælges median$\geq 0,5\%$, hvor fællesmængdernes 14 OTU'er dækker $61,46\%$.

Tabel 2

	max 0,1%	max 0,5%	max 1%
Reaktor T14	561	47	16
Reaktor T16	788	49	21
Reaktor T17	569	44	13
Foreningsmængden (alle reaktorerne)	885 (97,69%)	58(79,23%)	22 (66,94%)
Fællesmængden (alle reaktorerne)	451 (95,44%)	38(75,26%)	13 (59,52%)

Tabel 3 (Ændre caption ved alle tre tabeller)

	median $\geq 0,1\%$	median $\geq 0,5\%$	median $\geq 1\%$
Reaktor T14	193	16	5
Reaktor T16	197	18	5
Reaktor T17	207	19	6
Foreningsmængden	231 (91,28%)	20 (67,24%)	6 (49,28%)
Fællesmængden	169 (88,61%)	14 (61,46%)	5 (46,5%)

Tabel 4

	gennemsnit $\geq 0,1\%$	gennemsnit $\geq 0,5\%$	gennemsnit $\geq 1\%$
Reaktor T14	214	19	5
Reaktor T16	201	19	7
Reaktor T17	209	21	6
Foreningsmængden	238 (91,91%)	22 (68,79%)	7(51,40%)
Fællesmængden	172 (89,43%)	18 (65,89%)	5(46,5%)

2.4 Interpolation af ækvivalente tidspunkter

Ved analyse af en tidsserie er det nødvendigt, at observationerne er målt til ækvidistante tider. Da dette ikke er tilfældet for prøverne taget i biogasanlægget interpoleres observationer til ækvidistante tider. Dette gøres ved hjælp af en spline se afsnit 1.2. For at vurdere graden af spline, er forskellige grader af spline fittet for en udvalgt OTU i reaktor $T14$. Disse sammenlignes og den bedste fit udvælges til en kubisk spline. Derefter reguleres på antallet af knuder på samme måde og til sidst placeringen af knuder. Dette resulterer i to mulige spline, som er en kubisk spline med 15 knuder placeret til ækvidistante tider og en kubisk spline med 15 knuder placeret ved hjælp af percentiler. For at vurdere, hvilke af de to mulige placeringer af knuderne, som passer bedst til datasættet, er der foretaget kubisk spline med de to mulige placeringer af knuder for alle OTU'erne, se appendiks C. Ved en sammenligning af disse vurderes det, at knuderne placeret ved percentilerne giver den bedste fit på punkterne. Antallet af knuder for reaktor $T16$ og $T17$ er valgt til henholdsvis 12 og 13. Dette skyldes, at det rensede datasæt består af 23 og 26 observerede tider for henholdsvis $T15$ og $T16$.

2.5 Modellering

Modelleringen påbegyndes med en vektorautoregressiv model af orden en, da modellen anses for at være den simpleste multivariate tidsserie. Lasso straffen pålægges for at gøre modellen mindre kompleks og dermed undgå et overfit af datasættet. VAR(1) er konstrueret for alle reaktorer for at se, om der er en markant forskel på dynamikken af OTU'er imellem reaktorerne. VAR(1) viser, at der er en forskel på dynamikken af OTU'er imellem reaktorerne, idet overgangsmatricerne har ret forskellig struktur. PCA og et plot af tidsserierne viser dog, at der er en tendens til at de følger den samme dynamik. Idet formålet med projektet er at konstruere en model for abundansen af OTU'er henover tiden vil der ikke blive lagt yderligere vægt på forkellen imellem reaktorerne. Da der er flest observationer fra reaktor $T14$ vil modelleringen af OTU'erne blive udført baseret på reaktor $T14$. Efter at have estimeret VAR(1)-strukturen for T14 ønskes et bedre fit, hvorfor de univariate tidserier analyseres hver for sig med udgangspunkt i den fastlagte struktur fra A matricen for VAR(1). Viser A matricen eksempelvis, at abundansen af OTU 26 er korreleret med abundansen af OTU 26, OTU 12 og OTU 5 til en tidsenhed tilbage konstrueres en autoregresiv model af orden 1 med to forklarende variable, der angiver abundansen af henholdsvis OTU 12 og OTU 5 til en tidsenhed tilbage. På denne måde opnås et bedre estimat af regressionskoefficienterne, idet der ikke påføres en straf. Ved hjælp af t-test undersøges om estimaterne er signifikante. Er der et estimat, som ikke er signifikant, konstrueres en ny model uden den tilhørende variabel, hvorefter estimaterne testes igen. Til dette anvendes den implementerede funktion `modellering()`, se appendiks D. Viser estimaterne sig at være signifikante undersøges om residualprocessen opfylder stationaritet, ikke er serielt korreleret, og om residualerne opfylder normalitet.

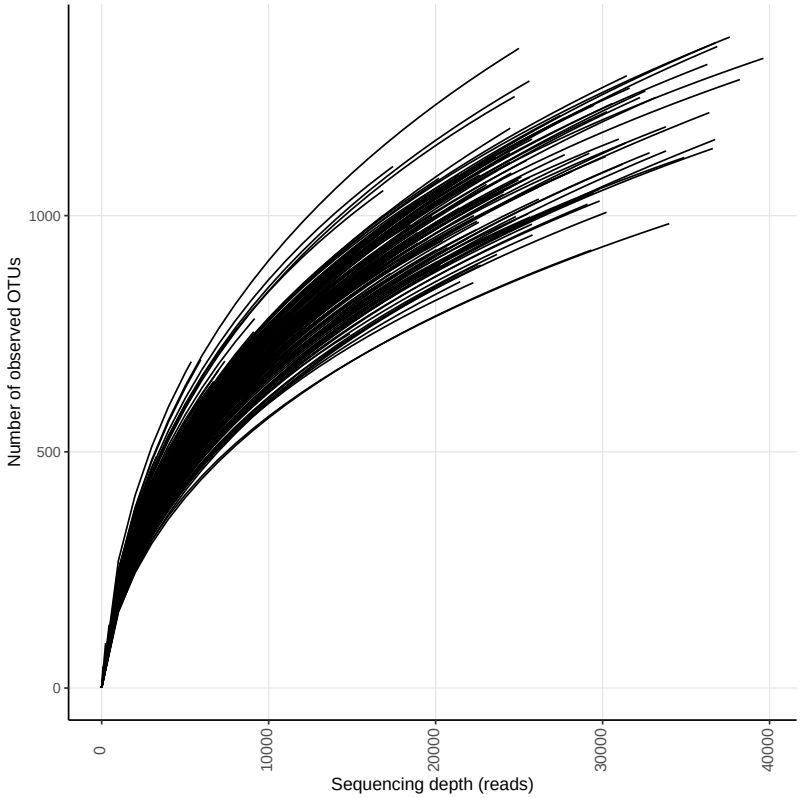
2.6 Modelvalidering

Størrelsen af lasso straffen, der anvendes til estimering af VAR(1), bestemmes ved at krydsvalidere en række sparse VAR(1) modeller. Modellen, med den tilhørende straf, som giver den laveste krydsvalidering, CV_{10}, er anvendt. Under krydsvalidering er $k = 10$ valgt for både at korrigere for at CV_{10} ikke er et overfit og at det skal være repræsentativt, se afsnit 1.6. I programmet **R** anvendes funktionen `fitVAR` fra pakken `sparsevar` til at finde modellen, og den tilhørende lassostraf, som har den mindste CV_{10}. For $k = 10$ vil der være mange mulige kombinationer, hvorpå observationerne kan inddeles i, hvilket kan resultere i

mange forskellige estimater af CV_{10}. Haves et tilstrækkeligt antal observationer således, at observationssættet er repræsentativt for den dynamik, der er i biogasanlægget, vil de mange estimater af CV_{10} ligge tæt på hinanden. Er dette ikke tilfældet, er det ofte en indikation af, at der er for få observationer, hvilket kan resultere i store forskelle i estimaterne af CV_{10}. Konsekvensen er, at hver gang `fitVAR` anvendes vil det resultere i forskellige sparse VAR(1). Af kapitel 3 observeres, at dette er tilfældet. Derfor ses på strukturen af de fittede sparse VAR(1) og modellen med flest ikke-nul indgange i diagonalen vælges, idet det virker naturligt med én autoregressiv komponent. For de OTU'er, hvor VAR(1) viser, at den ikke afhænger af én autoregressiv komponent, vurderes ud fra en sammenligning af PACF og ACF korrelogrammer, muligheden for at abundansen kan forklares som en AR(1).

3 Resultater

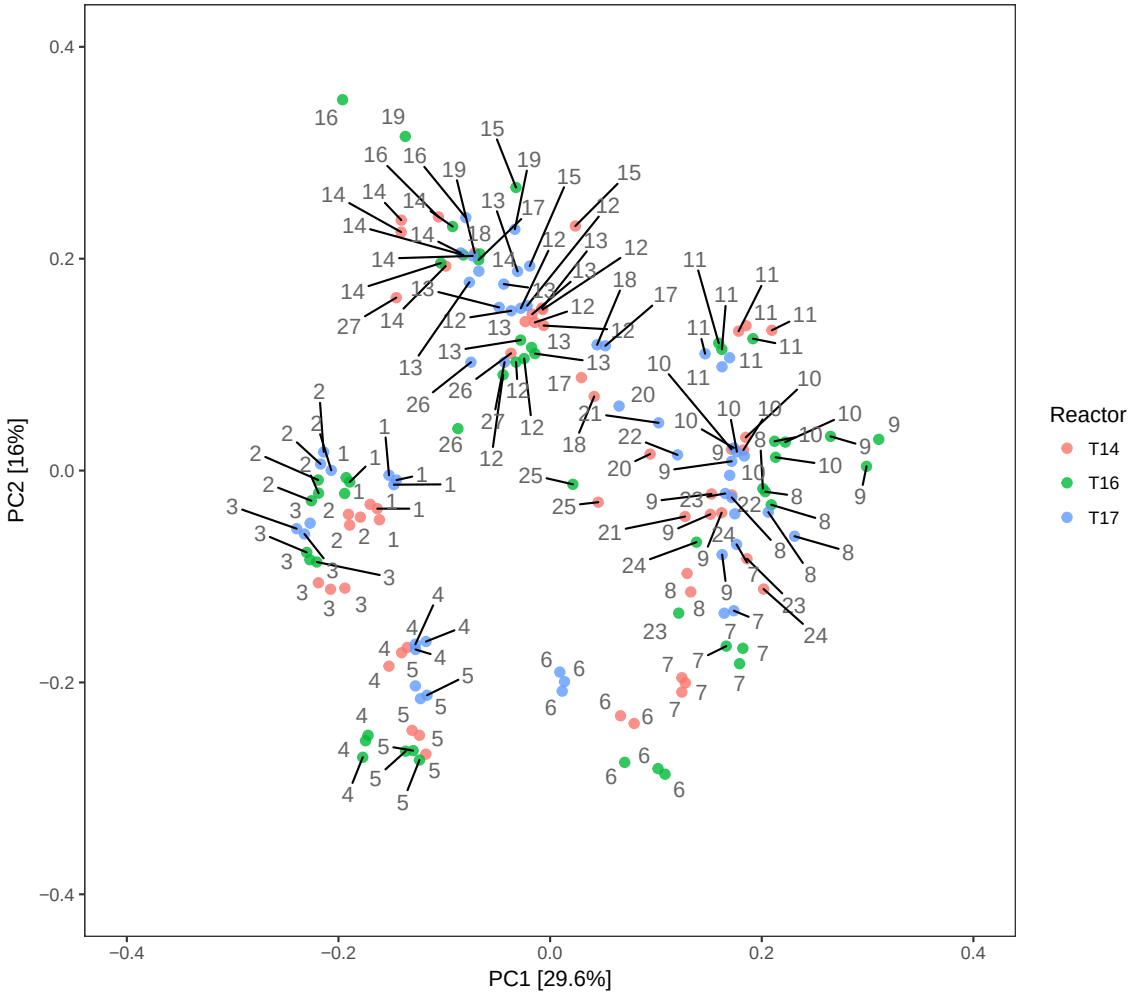
På figur 7 ses et plot med en rarefactionkurve for hver stikprøve. Prøver med en sekvensdybde på mindre end 3000 reads er fjernet, hvorved prøverne reduceres fra 174 til 161, OTU'er reduceres fra 3344 til 3343, og antallet af tilbageværende tidsobservationer er 27, 23, 26 for henholdsvis reaktor T14, T16 og T17. To af de fjernede prøver stammer fra prøvetagningsdag 7 taget i T14, se A.3. Af `Sample_id` fremgår det, at prøverne er sekventeret under samme omgang. Derfor fjernes den tilbageværende prøve fra samme omgang således, at der kun er et triplikat tilbage. Det samme gør sig gældende for to af prøverne fra prøvetagningsdag 12 taget i reaktor $T17$, hvorfor den tilbageværende prøve fra samme omgang fjernes.



Figur 7: Et plot der viser rarefactionkurver taget for hver stikprøve.

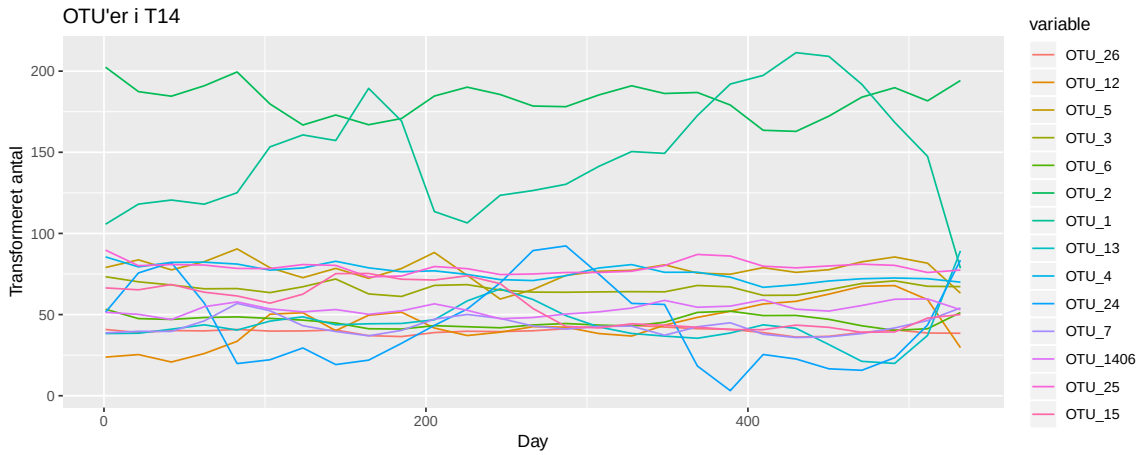
Af PCA plottet, se figur 8, er der en tendens til, at prøverne placerer sig i et mønster afhængig af tiden. Derudover ses en tendens til, at triplikaterne klynger sig sammen, hvilket indikerer, at der ikke er en markant forskel på prøverne indenfor en triplikat. Det samme kan konkluderes af heatmaps, se appendiks B. Ved at følge proceduren for udvælgelse af triplikater beskrevet i afsnit 2.2 beholdes prøven med flest antal reads pr. triplikat. Dette

resulterer i, at antallet af OTU'er reduceres til 3288.

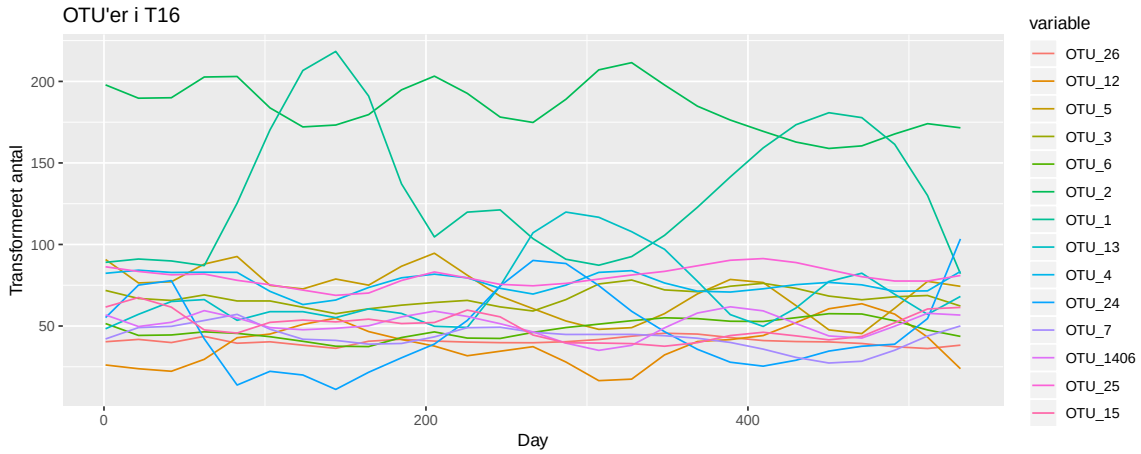


Figur 8: PCA plot over stikprøverne, der er tilbage efter fortyndingen.

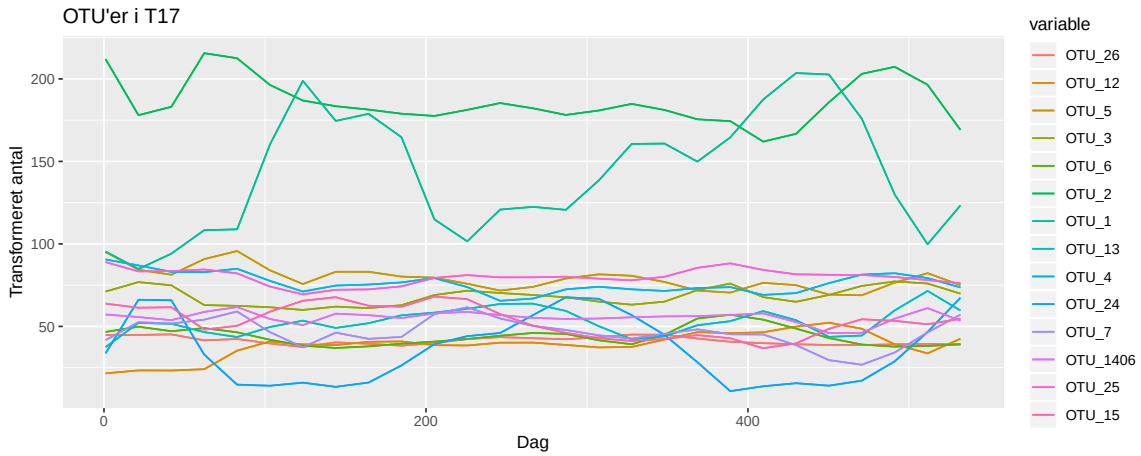
Herefter reduceres OTU'er yderligere ved hjælp af metoden beskrevet i afsnit 2.3, således at der til sidst er 14 OTU'er tilbage, se output 3.1. Figur 9, 10 og 11 viser tidsserierne af abundansen for de 14 OTU'er i hver reaktor. For alle tre reaktorer er det abundansen af OTU 2, der forekommer i høje koncentrationer over hele perioden. Ydermere ser det ud til at abundansen af OTU 2 påvirkes af abundansen af OTU 1, da en stigning/reducering af OTU 1 giver en reducering/stigning i abundansen af OTU 2. For OTU 1 ser det ud til at abundansen er bestemt af både OTU 2 og OTU 24.
Sammenlignes tidsserierne for de tre reaktorer ses en tendens til, at dynamikken imellem OTU'erne er ens imellem de tre reaktorer, dog viser estimaterne for sparse VAR(1), se appendiks E, at dette ikke er tilfældet.



Figur 9: Tidsserier for en delmængde af OTU'erne taget i reaktor T14.



Figur 10: Tidsserier for en delmængde af OTU'erne taget i reaktor T16.



Figur 11: Tidsserier for en delmængde af OTU'erne taget i reaktor T17.

```
             Kingdom             Phylum               Class                  Order
OTU_26   k__Bacteria     p__Firmicutes       c__Clostridia      o__Clostridiales
OTU_12   k__Bacteria
OTU_5    k__Bacteria     p__Firmicutes       c__Clostridia      o__Clostridiales
OTU_3    k__Bacteria     p__Firmicutes c__Erysipelotrichia o__Erysipelotrichales
OTU_6    k__Bacteria     p__Firmicutes       c__Clostridia      o__Clostridiales
OTU_2    k__Bacteria     p__Firmicutes       c__Clostridia      o__Clostridiales
OTU_1    k__Bacteria p__Actinobacteria   c__Actinobacteria    o__Actinomycetales
OTU_13   k__Bacteria     p__Firmicutes            c__OPB54  o__Hydrogenisporales
OTU_4    k__Bacteria     p__Firmicutes       c__Clostridia      o__Clostridiales
OTU_24   k__Bacteria     p__Firmicutes          c__Bacilli    o__Lactobacillales
OTU_7    k__Bacteria     p__Firmicutes       c__Clostridia      o__Clostridiales
OTU_1406 k__Bacteria     p__Firmicutes       c__Clostridia      o__Clostridiales
OTU_25   k__Bacteria     p__Firmicutes       c__Clostridia      o__Clostridiales
OTU_15   k__Bacteria     p__Firmicutes       c__Clostridia      o__Clostridiales
                          Family                          Genus Species
OTU_26        f__Clostridiaceae 1 g__Clostridium sensu stricto 1     s__
OTU_12
OTU_5         f__Clostridiaceae 1 g__Clostridium sensu stricto 1     s__
OTU_3       f__Erysipelotrichaceae                g__Turicibacter     s__
OTU_6     f__Peptostreptococcaceae
OTU_2         f__Clostridiaceae 1 g__Clostridium sensu stricto 1     s__
OTU_1         f__Actinomycetaceae                  g__Trueperella     s__
OTU_13                  f__MBA03                              g__     s__
OTU_4     f__Peptostreptococcaceae             g__Terrisporobacter     s__
OTU_24        f__Lactobacillaceae                g__Lactobacillus     s__
OTU_7     f__Caldicoprobacteraceae             g__Caldicoprobacter     s__
OTU_1406      f__Clostridiaceae 1 g__Clostridium sensu stricto 1     s__
OTU_25    f__Peptostreptococcaceae
OTU_15         f__Ruminococcaceae               g__Fastidiosipila     s__
```

Listing 3.1: Liste over de 14 OTU'er, der er tilbage efter rensning.

I nedenstående output ses en række estimater af lassostrafen (tuningsparameteren) baseret på krydsvalidering med $k = 10$, foretaget 40 gange. Som det ses i outputtet varierer størrelsen for lasso meget, hvilket indikerer, at der er for få observationer. Lassostraffen på $0,22$ er anvendt til estimering af sparse VAR(1) for reaktor T14.

```
lampda=vector()
diag=vector()
for(x in 1:40){set.seed(x)
  T14_fit=fitVAR(T14_data_matrix, p=1)
tmp=T14_fit$lambda
tmp1=sum(diag(T14_fit$A[[1]])==0)
lampda=c(lampda,tmp[1])
diag=c(diag,tmp1[1])}
lampda; diag
```

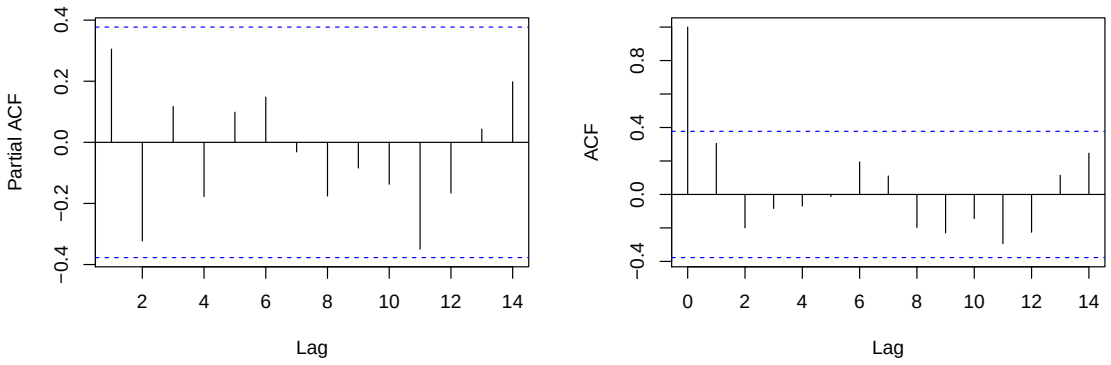
```
 [1] 0.7380110 0.5086824 0.3847996 0.6724482 0.4634924 0.4223170 0.5086824 0.5582783
 [9] 0.3847996 0.4223170 0.2201965 0.6127097 0.4634924 0.5086824 0.6127097 0.4634924
[17] 0.3847996 0.5086824 0.4223170 0.4223170 0.5086824 0.4223170 0.5582783 0.4634924
[25] 0.2652274 0.5582783 0.5582783 0.3847996 0.4223170 0.3847996 0.6127097 0.4223170
[33] 0.3847996 0.3847996 0.6127097 0.4634924 0.5086824 0.6127097 0.5582783 0.0412608
 [1] 6 5 4 6 5 5 5 5 4 5 2 6 5 5 6 5 4 5 5 5 5 5 5 5 2 5 5 4 5 4 6 5 4 4 6 5 5 6 5 3
```

Ved brug af denne lassostraf fittes en sparse VAR(1), hvorved nedenstående estimater for overgangsmatricen og interceptet er opnået. Ud fra overgangsmatricen er abundansen af hver OTU, med undtagelse af OTU 5 og OTU 1406, korreleret med sin egen abundans til lag en. PACF korrelogrammet for OTU 1406 viser dog, at der er en signifikant PACF til lag 1, se figur 13. PACF korrelogrammet for OTU 5 viser, at der en PACF for lag 1, men at denne ikke er signifikant, se figur 12.

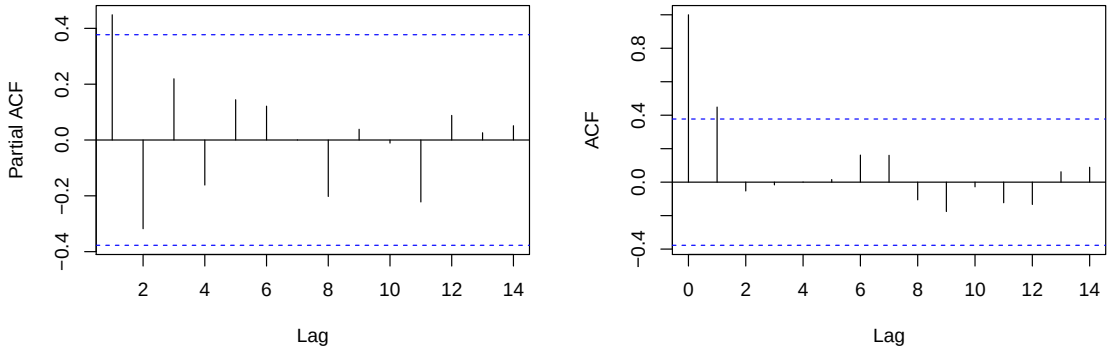
```
             OTU_26      OTU_12      OTU_5       OTU_3      OTU_6       OTU_2
OTU_26     0.1776413  0.00000000  0.0000000  0.0000000  0.0000000  0.00000000
OTU_12     0.1857081  0.07183886  0.0000000  0.00000000  0.0000000  0.00000000
OTU_5      0.0000000  0.00000000  0.0000000 -0.31033306  0.1400266  0.00000000
OTU_3      0.0000000  0.00000000  0.0000000  0.06781112  0.0000000  0.00000000
OTU_6      0.0000000  0.00000000  0.0000000  0.00000000  0.3635426  0.00000000
OTU_2     -0.6127198  0.00000000  0.2230034  0.00000000 -0.3957002  0.34330169
OTU_1      6.1629206 -0.07633670  0.0000000  0.00000000  2.6034070 -0.01117851
OTU_13    -0.7174985  0.00000000  0.0000000  0.39769726 -0.8039343  0.00000000
OTU_4      0.0000000  0.00000000  0.0000000  0.00000000  0.0000000  0.00000000
OTU_24    -1.6577241  0.00000000 -0.6495595  2.33256384 -0.5167425  0.29516546
OTU_7      0.0000000  0.00000000  0.1810734  0.00000000 -0.1449446  0.00000000
OTU_1406   0.0000000  0.00000000  0.0000000  0.00000000  0.0000000  0.00000000
OTU_25     0.0000000  0.00000000  0.0000000  0.00000000  0.1468354  0.00000000
OTU_15     0.0000000  0.00000000  0.2641461  0.15056230  0.0000000 -0.12192921
```

```
                OTU_1        OTU_13       OTU_4        OTU_24       OTU_7     OTU_1406
OTU_26     0.000000000  0.000000000  0.0000000  0.005394777  0.0000000  0.00000000
OTU_12     0.139238454  0.000000000 -0.4622357 -0.155189862  0.0000000  0.00000000
OTU_5      0.000000000 -0.246296743  0.3518191  0.000000000 -0.2196021  0.00000000
OTU_3      0.000000000 -0.010041509  0.0000000  0.000000000  0.0000000  0.00000000
OTU_6      0.000000000  0.000000000  0.0000000  0.000000000  0.0000000  0.05562825
OTU_2      0.000000000 -0.145974957  0.0000000  0.166161335  0.0000000  0.00000000
OTU_1      0.899602280  0.754723597  0.0000000 -0.167051664 -0.4752555  0.00000000
OTU_13    -0.055886175  0.557346492  0.0000000  0.000000000  0.5891046  0.97059421
OTU_4      0.000000000  0.000000000  0.5390880  0.022252112  0.0000000 -0.04610656
OTU_24    -0.209938017  0.000000000 -1.1060015  0.494289560  0.0000000  0.00000000
OTU_7     -0.028699376  0.000000000  0.0000000  0.000000000  0.3450238  0.00000000
OTU_1406   0.001039081 -0.132846790  0.0000000  0.000000000  0.0000000  0.00000000
OTU_25     0.000000000 -0.005882827  0.0000000  0.000000000  0.0000000  0.00000000
OTU_15     0.000000000  0.000000000  0.2144117  0.000000000  0.0000000  0.00000000
               OTU_25       OTU_15
OTU_26     0.0000000  0.00000000
OTU_12     0.0000000 -0.05952616
OTU_5      0.0000000  0.00000000
OTU_3      0.0000000  0.00000000
OTU_6      0.0000000  0.00000000
OTU_2      0.0000000  0.00000000
OTU_1      0.0000000  0.00000000
OTU_13     0.0000000  0.02560487
OTU_4      0.0000000  0.00000000
OTU_24    -1.2632757  0.00000000
OTU_7      0.0000000  0.00000000
OTU_1406   0.0000000 -0.02766429
OTU_25     0.0831477 -0.02161781
OTU_15     0.0000000  0.81059709
```

```
           intercept
OTU_26     -32.43836
OTU_12     -58.21866
OTU_5      -84.97144
OTU_3      -62.24781
OTU_6      -26.22737
OTU_2     -143.86263
OTU_1      349.17365
OTU_13      11.90408
OTU_4      -36.47506
OTU_24    -169.21173
OTU_7      -25.15811
OTU_1406   -60.67165
OTU_25     -67.34817
OTU_15      13.99800
```



Figur 12: PACF og ACF korrelogrammer for OTU 5 i reaktor T14.



Figur 13: PACF og ACF korrelogrammer for OTU 1406 i reaktor T14.

I nedenstående output ses, at den absolutte værdi af egenværdierne er mindre end en, hvorfor tidsserien er stationær, hvis residualerne også opfylder betingelserne for stationaritet.

```
 [1] 0.82709597 0.82709597 0.56149084 0.56149084 0.55160088 0.55160088
 [7] 0.38433461 0.25127846 0.15155305 0.15155305 0.08794015 0.06954806
[13] 0.06017344 0.06017344
```

Nedenfor er p-værdierne for Dickey-Fuller testene for residualprocessen vist, hvor det observeres at kun 5 ud af 14 tidsserier er stationære.

```
    OTU_26     OTU_12      OTU_5      OTU_3      OTU_6      OTU_2      OTU_1
0.01866254 0.99000000 0.01000000 0.33229609 0.02059865 0.01000000 0.33434325
    OTU_13      OTU_4     OTU_24      OTU_7   OTU_1406     OTU_25     OTU_15
0.20687446 0.17172548 0.04446006 0.13267261 0.38646747 0.27650989 0.18087951
```

P-værdien for OTU 12 er ekstrem høj, hvilket virker mærkværdig, da p-værdierne for de øvrige tidsserier ikke viser samme tendens. Der er derfor udført en Dickey-Fuller test for OTU 12 igen, men hvor én af residualerne er fjernet.

```
  Augmented Dickey-Fuller Test
data:  res[-c(26), 2]
Dickey-Fuller = -2.4823, Lag order = 2, p-value = 0.3886
alternative hypothesis: stationary
```

P-værdien falder fra $0,99$ til $0,39$ når en af observationerne er fjernet. Dette er et kæmpe fald, hvilket indikerer, at der er for få observationer og at man derfor skal forholde sig kritisk overfor troværdigheden af Dickey-Fuller testen.
I appendiks E.1 er der foretaget en residual analyse for at undersøge, hvor valid den estimerede sparse VAR(1) er. Af analysen fremgår det, at residualerne er tilnærmelsesvis normalfordelte, se figur 39 og 40. Ydermere observeres ud fra PACF korrelogrammerne for residualerne, se figur 34, at en mulig model for OTU 24 er en AR(3), idet der forekommer en signifikant korrelation for lag tre. Estimaterne af overgangsmatricen og interceptet er bestemt ved hjælp af funktionen `modellering()` til følgende:

```
             OTU_26    OTU_12    OTU_5     OTU_3     OTU_6     OTU_2       OTU_1
OTU_26    0.6480416 0.0000000  0.00000 0.0000000 0.0000000 0.0000000  0.00000000
OTU_12    0.0000000 0.5373823  0.00000 0.0000000 0.0000000 0.0000000 -0.22180528
OTU_3     0.0000000 0.0000000  0.00000 0.4058137 0.0000000 0.0000000  0.00000000
OTU_6     0.0000000 0.0000000  0.00000 0.0000000 0.6851194 0.0000000  0.00000000
OTU_2     0.0000000 0.0000000  0.00000 0.0000000 0.0000000 0.4788225  0.00000000
OTU_1     0.0000000 0.0000000  0.00000 0.0000000 0.0000000 0.0000000  0.78090016
OTU_13    0.0000000 0.0000000  0.00000 0.0000000 0.0000000 0.0000000  0.00000000
OTU_4     0.0000000 0.0000000  0.00000 0.0000000 0.0000000 0.0000000  0.00000000
OTU_24    0.0000000 0.0000000 -1.12382 0.0000000 0.0000000 0.0000000 -0.56750049
OTU_7     0.0000000 0.0000000  0.00000 0.0000000 0.0000000 0.0000000 -0.09685152
OTU_1406  0.0000000 0.0000000  0.00000 0.0000000 0.0000000 0.0000000  0.00000000
```

```
OTU_25    0.0000000 0.0000000  0.00000 0.0000000 0.0000000 0.0000000  0.00000000
OTU_15    0.0000000 0.0000000  0.00000 0.0000000 0.0000000 0.0000000  0.00000000
             OTU_13     OTU_4      OTU_24      OTU_7  OTU_1406     OTU_25
OTU_26    0.0000000 0.0000000  0.00000000  0.0000000 0.0000000  0.0000000
OTU_12    0.0000000 0.0000000  0.00000000  0.0000000 0.0000000  0.0000000
OTU_3     0.0000000 0.0000000  0.00000000  0.0000000 0.0000000  0.0000000
OTU_6     0.0000000 0.0000000  0.00000000  0.0000000 0.0000000  0.0000000
OTU_2     0.0000000 0.0000000  0.18504926  0.0000000 0.0000000  0.0000000
OTU_1     0.0000000 0.0000000 -0.84546782 -2.3141986 0.0000000  0.0000000
OTU_13    0.5176783 0.0000000  0.00000000  1.3218567 0.0000000  0.0000000
OTU_4     0.0000000 0.8108694  0.07778483  0.0000000 0.0000000  0.0000000
OTU_24    0.0000000 0.0000000  0.71748581  0.0000000 0.0000000 -2.3673297
OTU_7     0.0000000 0.0000000  0.00000000  0.5942318 0.0000000  0.0000000
OTU_1406  0.0000000 0.0000000  0.00000000  0.0000000 0.4389289  0.0000000
OTU_25    0.0000000 0.0000000  0.00000000  0.0000000 0.0000000  0.5700615
OTU_15    0.0000000 0.0000000  0.00000000  0.0000000 0.0000000  0.0000000
            OTU_15
OTU_26    0.000000
OTU_12    0.000000
OTU_3     0.000000
OTU_6     0.000000
OTU_2     0.000000
OTU_1     0.000000
OTU_13    0.000000
OTU_4     0.000000
OTU_24    0.000000
OTU_7     0.000000
OTU_1406  0.000000
OTU_25    0.000000
OTU_15    0.912014
```

```
          Intercept
OTU_26     39.72757
OTU_12     26.78978
OTU_3      66.49488
OTU_6      46.77026
OTU_2     173.91861
OTU_1     275.31615
OTU_13    -10.77505
OTU_4      71.91952
OTU_24    410.76871
OTU_7      58.08169
OTU_1406   53.39862
OTU_25     79.60797
OTU_15     48.39052
```

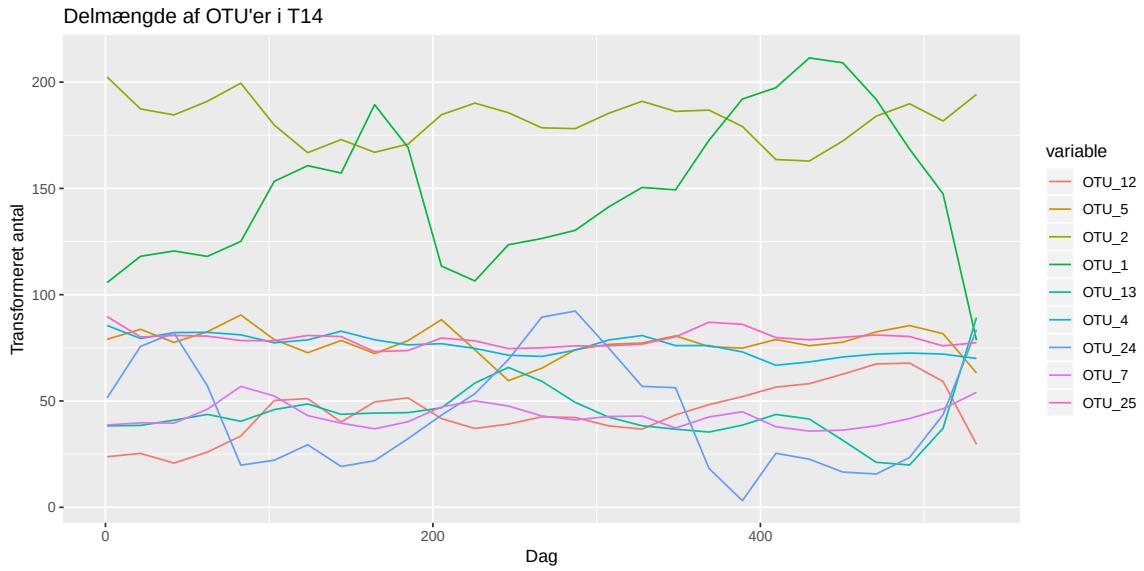
Nedenfor ses p-værdier for Dickey-Fuller testene udført på residualprocesserne, hvor det observeres, at 5 ud af 13 tidsserier er stationære. Udføres Dickey-Fuller testene, hvor man har fjernet nogle af residualer sker der et fald i p-værdierne, se appendiks F. Det viser igen, at der er få observationer, hvorfor man skal forholde sig kritisk til troværdigheden af Dickey-Fuller testen.

```
    OTU_26     OTU_12      OTU_3      OTU_6      OTU_2      OTU_1     OTU_13      OTU_4
0.04449574 0.22402319 0.01000000 0.44467203 0.44467203 0.06639798 0.95224206 0.06869618
OTU_24      OTU_7   OTU_1406     OTU_25     OTU_15
0.01000000 0.07785184 0.01758358 0.01773627 0.20818267
```

Nedenfor er p-værdierne for Box-Pierce testen vist for residualprocesserne. Nulhypotesen i Box-Pierce testen er, at ACF er lig nul for alle lags. Testene viser derfor, at man ikke kan forkaste nulhypotesen, om at der ingen seriel korrelation er.

```
    OTU_26     OTU_12      OTU_3      OTU_6      OTU_2      OTU_1     OTU_13
0.16883189 0.55252848 0.72376706 0.22025038 0.22025038 0.36848668 0.30483362
     OTU_4      OTU_24      OTU_7   OTU_1406     OTU_25     OTU_15
0.13961251 0.42549839 0.18303299 0.49863124 0.38529677 0.02993952
```

I appendiks F er ACF og PACF korrelogrammerne for residualerne vist for alle de univariate tidsserier. Det observeres, at OTU 3, OTU 7 og OTU 25 har en signifikant PACF for lag to. Dette indikerer, at en ARX(2) vil passe bedre til modellen, se eksempel 5. I appendiks F er der foretaget en residualanalyse af de univariate tidsserier, hvor det viser sig, at residualprocesserne er normalfordelte. Ved fjernelse af OTU'erne som kun afhænger af sig selv og som ikke har indflydelse på en af de øvrige OTU'er opnås ialt 9 OTU'er. Tidsserierne for de 9 OTU'er ses på figur 14.



Figur 14: Tidsserier af OTU'erne, der indgår i ARX(1).

For Lotka-Volterra modellen er følgende M matrix og alfa bestemt:

	OTU_12	OTU_5	OTU_2	OTU_1	OTU_13
OTU_12	-0.01045202	0.00000000	0.000000000	-0.005011293	0.00000000
OTU_5	0.00000000	-0.01294219	0.000000000	0.000000000	0.00000000
OTU_2	0.00000000	0.00000000	-0.002862456	0.000000000	0.00000000
OTU_1	0.00000000	0.00000000	0.000000000	-0.001468864	0.00000000
OTU_13	0.00000000	0.00000000	0.000000000	0.000000000	-0.01099229
OTU_4	0.00000000	0.00000000	0.000000000	0.000000000	0.00000000
OTU_24	0.00000000	-0.02529556	0.000000000	-0.012773611	0.00000000
OTU_7	0.00000000	0.00000000	0.000000000	-0.002248058	0.00000000
OTU_25	0.00000000	0.00000000	0.000000000	0.000000000	0.00000000

	OTU_4	OTU_24	OTU_7	OTU_25
OTU_12	0.000000000	0.000000000	0.000000000	0.000000000
OTU_5	0.000000000	0.000000000	0.000000000	0.000000000
OTU_2	0.000000000	0.001016343	0.000000000	0.000000000
OTU_1	0.000000000	-0.005668089	-0.015514587	0.000000000
OTU_13	0.000000000	0.000000000	0.030125597	0.000000000
OTU_4	-0.002490753	0.001024387	0.000000000	0.000000000
OTU_24	0.000000000	-0.006358984	0.000000000	-0.053285150
OTU_7	0.000000000	0.000000000	-0.009418441	0.000000000
OTU_25	0.000000000	0.000000000	0.000000000	-0.005427264

	alfa
OTU_12	0.6052672
OTU_5	0.0000000
OTU_2	0.9552106
OTU_1	1.8457432
OTU_13	-0.2455674
OTU_4	0.9471434
OTU_24	9.2458066
OTU_7	1.3481566
OTU_25	1.0049192

$\boldsymbol{\alpha}$-estimatet for Lotka-Volterra modellen viser, at vækshastighederne for alle OTU'er med undtagelse af OTU 5 og OTU 13 er positive. Ydermere er vækshastigheden for OTU 24 mindst 4 gange så stor som for de øvrige OTU'er. For OTU 5 er vækshastigheden nul, imens den er negativ for OTU 13. Alle diagonalelementerne i interaktionsmatricen er negative. Dette betyder, at OTU'erne vil være i stand til at opnå sin bærekapacitet uden tilstedeværelse af de øvrige OTU'er [30]. Abundansen af alle OTU'erne med undtagelse af OTU 5 bliver påvirket af mindst en anden OTU. Derimod har OTU 5 en negativ indvirkning på OTU 24. Forholdet imellem OTU 5 og OTU 24 kaldes ammenalisme [35]. Af interaktionsmatricen observeres det, at kun to af OTU'erne, OTU 24 og OTU 7, har en positiv virkning på henholdsvis OTU 2, OTU 4 og OTU 13. Størrelsen på interaktionerne er generelt ret lave, hvilke indikere at der ikke er en stærk korrelation mellem OTU'erne.

4 Konklusion

Formålet med specialet er, at beskrive den tidsafhængige mikrobielle dynamik i et biogasanlæg. Dette er gjort ved at bruge OTU'er, som en enhed for mikroorganismerne. I begyndelsen af specialeforløbet blev der udleveret et datasæt for både mikroorganismer fra bakterier og arkaer, hvor det blev observeret, at biogasanlægget indeholder flere forgreninger (32) af bakterier og en bestemt forgrening, kaldet metanogener, af arkaer. I projektet er der kun fokuseret på det ene datasæt indeholdende bakteriemålinger. Dette skyldes, at bakterier og arkaer har forskellige funktioner, hvor bakterier påbegynder den anaerobe udrådning, igennem tre processer, og ender med tre slutprodukterne (acetat, karbondioxid og hydorxid) som metanogener omdanner til biogas.
I projektet er der estimeret forskellige modeller, herunder sparse VAR(1) og VARX(1), for den tidsmæssige dynamik mellem 14 OTU'er ud af 3288 OTU'er. Af de 14 OTU'er er det kun 9 af OTU'erne, som har en indvirkning på hinanden. Dynamikken mellem de 9 OTU'er er også beskrevet ved hjælp af en Lotka-Volterra model. Af Lotka-Volterra modellen konkluderes, at forholdet mellem tre af OTU'erne kan beskrives som kommenalisme imens forholdet mellem de øvrige kan beskrives som konkurrence, ammenalisme eller ingen interaktion.
Ud fra de estimerede modeller kan det konkluderes, at der forekommer en (mindre) interaktion på tværs af OTU'erne. Dette vidner, om at dynamikken mellem OTU'erne er meget kompleks. Dermed støtter dette op, om at det mikrobielle samfund i et biogasanlæg er meget komplekst. Man bør derfor inddrage andre parametre, såsom substrat, til at forklare dynamikken mellem OTU'erne. Dette kan gøres ved at udvide den estimerede ARX(1) med eksempelvis substrat som en eksogen variabel.
Den estimerede VARX(1) model er baseret på prøveserien fra reaktor $T14$. Denne model er dog ikke reproducerbar i forhold til (at simulere abundansen af mikroorganismerne i) en af de øvrige reaktorer. Dette skyldes, at dynamikken mellem de tre reaktorer er forskellig fra hinanden på trods af at PCA og tidsserierne viser den modsatte tendens. Den estimerede VARX(1) er derfor heller ikke reproducerbar i forhold til et andet biogasanlæg. Flere af de statistiske analyser, herunder krydsvalidering af lasso-straffen, Dickey-Fuller test og Box-Juing test, viser at der er for få observationer. Dette reducerer troværdigheden af de estimerede modeller. At estimaterne for sparse VAR(1) er forskellige er (formegentlig) et resultat af manglende observationer.
Sammenfattende bidrager mit speciale til udviklingen af metoder til, at beskrive tidsdynamikken mellem OTU'erne i et biogasanlæg. Det vil dog kræve en del flere data at bidrage til en dybere forståelse af de biologiske mekanismer.

Bibliografi

[1] *NomiGas*. URL: https://www.en.bio.aau.dk/nomigas.

[2] *Dokumentationsdelen til Klimakommissionens samlede rapport GRØN ENERGI. VEJEN MOD ET DANSK ENERGISYSTEM UDEN FOSSILE BRÆNDS- LER*. Sep. 2010. ISBN: 978-87-7844-882-8. URL: https://www.klimabevaegelsen.dk/images/tema/viden/rapporter/klimakommissionendokumentation.pdf.

[3] Hannah Ritchie og Max Roser. "CO and other Greenhouse Gas Emissions". I: (2018). URL: https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions.

[4] Shahriar Shafiee og Erkan Topal. "When will fossil fuel reserves be diminished?" I: *Energy policy* 37.1 (2009), s. 181–189.

[5] *BP Statistical Review of World Energy 2018*. 67. udg. Jun. 2018. URL: https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2018-full-report.pdf.

[6] Amanda E. Olsen Daniel J. Caruna, red. *Anaerobic digestion : processes, products, and applications*. Nova Science Publishers, 2012.

[7] Jude Awele Okolie Kayode Feyisetan Adekunle. "A review of biochemical process of anaerobic digestion". I: *Advances in Bioscience and Biotechnology* (2015).

[8] Ying He Jing Fang Per Halkjaer Nielsen Willy Verstraete Nico Boon Jo De Vrieze Aaron Marc Saunders. "Ammonia and temperature determine potential clustering in the anaerobic digestion microbiome". I: *Water research* (2015).

[9] Mahesh Kumar Malav Shakeel A Khan Sandeep Kumar Lal Chand Malav. "Biogas slurry: Source of nutrients for eco-friendly agriculture". I: *International J. Ext. Res* (2015).

[10] Rikke Lybæk og Tyge Kjær. "Municipalities as facilitators, regulators and energy consumers for enhancing the dissemination of biogas technology in Denmark". I: *International Journal of Sustainable Energy Planning and Management* 8 (2015), s. 17–30.

[11] Peter Jacob Jørgensen. *Biogas – green energy*. 2. edition. Faculty of Agricultural Sciences, Aarhus University. ISBN: 978-87-992243-2-1.

[12] K.C. Surendra Shilva Shrestha Karthik Rajendran Hans Oechsner Li Xie Samir Kumar Khanal Chayanon Sawatdeenarunat Duc Nguyen. "Anaerobic biorefinery: current status, challenges, and opportunities". I: *Bioresource technology* (2016).

[13] Miljøudvalget 2011-2012. *Aftale om Grøn Vækst*. Jun. 2009. URL: https://www.ft.dk/samling/20111/almdel/miu/bilag/21/1030569.pdf.

[14] *Danske energipolitik 2012-2020*. URL: https://ens.dk/sites/ens.dk/files/Energibesparelser/aftale_22-03-2012_final_ren.doc.pdf.

[15] Energistyrelsen. *Dansk produktion af biogas*. URL: https://ens.dk/ansvarsomraader/bioenergi/produktion-af-biogas.

[16] Energistyrelsen. *Energistatistik 2017*. URL: https://ens.dk/sites/ens.dk/files/Statistik/pub2017dk.pdf.

[17] Dieter Deublein og Angelika Steinhauser. *Biogas from waste and renewable resources: an introduction*. Udg. af 2. edition. John Wiley & Sons, 2010. ISBN: 978-3-527-32798-0.

[18] *Maarbjerg energy center*. URL: https://www.maabjergenergycenter.dk/.

[19] Stuart H Hurlbert. "The nonconcept of species diversity: a critique and alternative parameters". I: *Ecology* (1971).

[20] David S. Matteson David Ruppert. *Statistics and data analysis for financial engineering with R examples*. Udg. af 2. udgave. Bd. 13. Springer, 2015.

[21] Daniel Zelterman. *Applied multivariate statistics with R*. Springer. ISBN: 978-3-319-14092-6.

[22] Ognyan Kounchev. *Multivariate polysplines: applications to numerical and wavelet analysis*. Academic Press, 2001.

[23] Gareth James m.fl. *An introduction to statistical learning*. Bd. 112. Springer, 2013. ISBN: 978-1-4614-7138-7.

[24] Jerome Friedman Trevor Hastie Robert Tibshirani. "The elements of statistical learning: Data mining, Inference, and Prediction". I: *International Statistical Review* (2009). ISSN: 978-0-387-84857-0.

[25] Kevin P. Murphy. *Machine learning, a probabilistic perspective*. Massachusetts Institute of Technology, 2012.

[26] URL: www.quora.com/Why-is-it-that-the-lasso-unlike-ridge-regression-results-in-coefficient-estimates-that-are-exactly-equal-to-zero.

[27] Allan C. Elliot Wayne A. Woodward Henry L. Grey. *Applied time series analysis, with R*. Udg. af 2. edition. 2017. ISBN: 978-1-4987-3422-6.

[28] David S. Stoer Robert H. Shumway. *Time Series Analysis and Its Applications*. 4. edition. ISBN: 978-3-319-52451-1.

[29] Silviu-Iulian Niculescu Arben Çela Xu-Guang Li Jun-Xiu Chen. "New insights in stability analysis of delayed Lotka–Volterra systems". I: *Journal of the Franklin Institute* (2018).

[30] Nora C. Toussaint Charlie G. Buffie Gunnar Rätsch Eric G. Pamer Chris Sander João B. Xavier Richard R. Stein Vanni Bucci. "Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota". I: (2013).

[31] Gregory C. Reinsel Greta M. Ljung George E. P. Box Gwilym M. Jenkins. *TIME SERIES ANALYSIS Forecasting and Control*. Udg. af 4. udgave. WILEY.

[32] Sam Behjati og Patrick S Tarpey. "What is next generation sequencing?" I: (2013).

[33] *sillva*. URL: https://www.arb-silva.de/.

[34] Søren Michael Karst Mads Albertsen Rasmus Hansen Kirkegaard Morten Simonsen Dueholm Per Halkjær Nielsen. "Molecular methods". I: *Molecular methods*. 2016.

[35] Guanglu Zhang m.fl. "System evolution prediction and manipulation using a Lotka–Volterra ecosystem model". I: *Design Studies* (2018).

A R-koder

Dette appendiks indeholder kodning foretaget i programmet R, hvilket gør, at læseren nemt kan fremkalde resultater opnået i projektet.

A.1 Eksempler

Datasættene `otu_bak_final.Rda` og `meta_fortyndet.Rda` konstrueret i afsnit A.3 er anvendt til eksempel 1.2. R-koderne anvendt er følgende:

```
load("otu_bak_final.Rda")
load("meta_fortyndet.Rda")
data_subset <- amp_load(otutable = otu_bak_final,metadata = meta_fortyndet)
pca=amp_ordinate(data_subset, sample_label_by = "Data_Point", sample_color_by = "Reactor")+ylim
    (-0.4,0.4)+xlim(-0.4,0.4) + geom_point(size=0.01)+
  theme_bw() +theme(aspect.ratio=1)+
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
pca_screeplot=amp_ordinate(data_subset, detailed_output = T)$screeplot+ylab("varians i procent"
    )+xlab("principielle komponenter")
amp_ordinate(data_subset, detailed_output = T)$dspecies
```

```
             OTU          PC1           PC2  Kingdom             Phylum
OTU_1       OTU_1  0.78367350 -0.139222216 Bacteria     Actinobacteria
OTU_24     OTU_24 -0.45547661 -0.307929845 Bacteria          Firmicutes
OTU_2       OTU_2 -0.25835846  0.131442171 Bacteria          Firmicutes
OTU_13     OTU_13 -0.17998350 -0.210422813 Bacteria          Firmicutes
OTU_12     OTU_12  0.16436934 -0.046198567 Bacteria k__Bacteria_OTU_12
OTU_5       OTU_5 -0.03656932  0.154611443 Bacteria          Firmicutes
OTU_4       OTU_4 -0.10498872  0.074607163 Bacteria          Firmicutes
OTU_7       OTU_7 -0.10021570  0.035609450 Bacteria          Firmicutes
OTU_15     OTU_15 -0.08261248  0.061026501 Bacteria          Firmicutes
OTU_25     OTU_25 -0.06783735  0.040713352 Bacteria          Firmicutes
OTU_3       OTU_3 -0.07807792  0.011654491 Bacteria          Firmicutes
OTU_1406 OTU_1406 -0.02176569  0.054903312 Bacteria          Firmicutes
OTU_26     OTU_26 -0.04073455  0.004093997 Bacteria          Firmicutes
OTU_6       OTU_6 -0.03843111  0.008824766 Bacteria          Firmicutes
                    Class              Order                Family
OTU_1       Actinobacteria     Actinomycetales      Actinomycetaceae
OTU_24             Bacilli     Lactobacillales      Lactobacillaceae
OTU_2          Clostridia       Clostridiales      Clostridiaceae 1
OTU_13               OPB54   Hydrogenisporales                MBA03
OTU_12   k__Bacteria_OTU_12 k__Bacteria_OTU_12    k__Bacteria_OTU_12
OTU_5          Clostridia       Clostridiales      Clostridiaceae 1
OTU_4          Clostridia       Clostridiales Peptostreptococcaceae
OTU_7          Clostridia       Clostridiales Caldicoprobacteraceae
OTU_15         Clostridia       Clostridiales       Ruminococcaceae
OTU_25         Clostridia       Clostridiales Peptostreptococcaceae
OTU_3      Erysipelotrichia Erysipelotrichales   Erysipelotrichaceae
OTU_1406       Clostridia       Clostridiales      Clostridiaceae 1
OTU_26         Clostridia       Clostridiales      Clostridiaceae 1
OTU_6          Clostridia       Clostridiales Peptostreptococcaceae
                              Genus                                  Species
OTU_1                     Trueperella                    g__Trueperella_OTU_1
OTU_24                   Lactobacillus                  g__Lactobacillus_OTU_24
OTU_2      Clostridium sensu stricto 1    g__Clostridium sensu stricto 1_OTU_2
OTU_13                 f__MBA03_OTU_13                        f__MBA03_OTU_13
OTU_12               k__Bacteria_OTU_12                     k__Bacteria_OTU_12
OTU_5      Clostridium sensu stricto 1    g__Clostridium sensu stricto 1_OTU_5
OTU_4                  Terrisporobacter              g__Terrisporobacter_OTU_4
OTU_7                  Caldicoprobacter              g__Caldicoprobacter_OTU_7
OTU_15                   Fastidiosipila                g__Fastidiosipila_OTU_15
OTU_25   f__Peptostreptococcaceae_OTU_25        f__Peptostreptococcaceae_OTU_25
OTU_3                     Turicibacter                   g__Turicibacter_OTU_3
OTU_1406   Clostridium sensu stricto 1 g__Clostridium sensu stricto 1_OTU_1406
OTU_26     Clostridium sensu stricto 1   g__Clostridium sensu stricto 1_OTU_26
```

```
OTU_6     f__Peptostreptococcaceae_OTU_6          f__Peptostreptococcaceae_OTU_6
               dist
OTU_1    0.633526982
OTU_24   0.302279728
OTU_2    0.084026140
OTU_13   0.076671820
OTU_12   0.029151589
OTU_5    0.025242013
OTU_4    0.016588860
OTU_7    0.011311220
OTU_15   0.010549056
OTU_25   0.006259483
OTU_3    0.006231988
OTU_1406 0.003488119
OTU_26   0.001676064
OTU_6    0.001554827
```

A.2 Uoverensstemmelser

Fejl i `Data_Point` rettes.

```
meta <-read.csv("metadataet.csv", sep = ";", header=TRUE,
                stringsAsFactors = FALSE)
meta <- meta[order(meta$Re,meta$Day),]
meta[c(which(meta$Data_Point==1),which(meta$Data_Point==2)),]
```

```
Seq_ID        Sample_ID Reactor Collection_Date Data_Point Week Day
2  16SAMP-9760 LIB-NDJ-0145-A-1      T14         20150625          1   25   1
3  16SAMP-9761 LIB-NDJ-0145-B-1      T14         20150625          1   25   1
4  16SAMP-9762 LIB-NDJ-0145-C-1      T14         20150625          1   25   1
40 16SAMP-9798 LIB-NDJ-0158-A-1      T16         20150625          1   25   1
41 16SAMP-9799 LIB-NDJ-0158-B-1      T16         20150625          1   25   1
42 16SAMP-9800 LIB-NDJ-0158-C-1      T16         20150625          1   25   1
43 16SAMP-9801 LIB-NDJ-0159-A-1      T16         20150709          1   28  15
79 16SAMP-9837 LIB-NDJ-0171-A-1      T17         20150625          1   25   1
80 16SAMP-9838 LIB-NDJ-0171-B-1      T17         20150625          1   25   1
81 16SAMP-9839 LIB-NDJ-0171-C-1      T17         20150625          1   25   1
82 16SAMP-9840 LIB-NDJ-0172-A-1      T17         20150709          1   28  15
5  16SAMP-9763 LIB-NDJ-0146-A-1      T14         20150709          2   28  15
6  16SAMP-9764 LIB-NDJ-0146-B-1      T14         20150709          2   28  15
7  16SAMP-9765 LIB-NDJ-0146-C-1      T14         20150709          2   28  15
44 16SAMP-9802 LIB-NDJ-0159-B-1      T16         20150709          2   28  15
45 16SAMP-9803 LIB-NDJ-0159-C-1      T16         20150709          2   28  15
83 16SAMP-9841 LIB-NDJ-0172-B-1      T17         20150709          2   28  15
84 16SAMP-9842 LIB-NDJ-0172-C-1      T17         20150709          2   28  15
```

```
meta$Data_Point[which(meta$Seq_ID=="16SAMP-9801")]="2"
meta$Data_Point[which(meta$Seq_ID=="16SAMP-9840")]="2"
```

Kontrolprøverne og de tilhørende OTU'er slettes fra datasættene.

```
meta[which(meta$Reactor=="Control"),]
```

```
 Seq_ID                Sample_ID Reactor Collection_Date Data_Point Week Day
87   16SAMP-9845     LIB-NDJ-POS-2016-MAA1 Control              NA       <NA>   NA  NA
119  16SAMP-9877     LIB-NDJ-POS-2016-MAA2 Control              NA       <NA>   NA  NA
177 16SAMP-17038 LIB-NDJ-POS-bV13-20170508 Control              NA       <NA>   NA  NA
```

```
meta=meta[-which(meta$Reactor=="Control"),]
otu_bak<- read.csv("otutabelet.csv", sep=";", header=TRUE,
                   stringsAsFactors=FALSE, check.names = FALSE)
Control=c("16SAMP-9845","16SAMP-9877","16SAMP-17038")
Control_OTU=vector()
for(i in 1:length(Control)){
  tmp=which(colnames(otu_bak)==Control[i])
  Control_OTU=c(Control_OTU,tmp[1])
}
otu_bak=subset(otu_bak,select = -Control_OTU)
otu_bak=otu_bak[apply(otu_bak[,2:175], 1, max)>0,]
```

Udsnit af data med prøver som er sekventeret over to omgange.

```
meta[c(intersect(which(meta$Data_Point==7), which(meta$Reactor=="T14")),
intersect(which(meta$Data_Point==12), which(meta$Reactor=="T17"))),]
```

```
   Seq_ID          Sample_ID Reactor Collection_Date Data_Point Week Day
19    16SAMP-9777    LIB-NDJ-0151-A-1      T14         20151008          7   41 106
20    16SAMP-9778    LIB-NDJ-0151-B-1      T14         20151008          7   41 106
21    16SAMP-9779    LIB-NDJ-0151-C-1      T14         20151008          7   41 106
120 16SAMP-12088 LIB-NDJ-0151-A-A-2      T14         20151008          7   41 106
121 16SAMP-12089 LIB-NDJ-0151-B-A-2      T14         20151008          7   41 106
122 16SAMP-12090 LIB-NDJ-0151-C-A-2      T14         20151008          7   41 106
113   16SAMP-9871    LIB-NDJ-0182-A-1      T17         20160114         12    2 204
114   16SAMP-9872    LIB-NDJ-0182-B-1      T17         20160114         12    2 204
115   16SAMP-9873    LIB-NDJ-0182-C-1      T17         20160114         12    2 204
123 16SAMP-12091 LIB-NDJ-0182-A-A-2      T17         20160114         12    2 204
124 16SAMP-12092 LIB-NDJ-0182-B-A-2      T17         20160114         12    2 204
125 16SAMP-12093 LIB-NDJ-0182-C-A-2      T17         20160114         12    2 204
```

A.3 Validation af sekvenser

Fjernelse af stikprøver der har et total reads på eller mindre end 3000 reads pr. stikprøve.

```
sumreads=vector()
for(x in 1:length(meta$Seq_ID)){
sumreadsen=colSums(otu_bak[,2:175])[meta$Seq_ID[x]]
sumreads=c(sumreads,sumreadsen)
  }
meta_tr=meta
meta_tr$sum_reads=sumreads
meta_fortyndet=meta[which(meta_tr$sum_reads>=3000),]
tmpp=vector()
for(x in 1:length(meta$Seq_ID[which(meta_tr$sum_reads<3000)])){
  tmp=which(colnames(otu_bak)==meta$Seq_ID[which(meta_tr$sum_reads<3000)][x])
  tmpp=c(tmpp,tmp)
}
otu_bak_fortyndet=otu_bak[,-tmpp]
otu_bak_fortyndet=otu_bak_fortyndet[apply(otu_bak_fortyndet[,2:162], 1, max)>0,]
```

Udsnit af de tilbageværende stikprøver og som er sekventeret over to omgange.

```
meta_fortyndet[c(intersect(which(meta_fortyndet$Data_Point==7), which(meta_fortyndet$Reactor==
    T14")), intersect(which(meta_fortyndet$Data_Point==12), which(meta_fortyndet$Reactor=="T17"
    ))),]
```

```
Seq_ID          Sample_ID Reactor Collection_Date Data_Point Week Day
20    16SAMP-9778    LIB-NDJ-0151-B-1      T14         20151008          7   41 106
120 16SAMP-12088 LIB-NDJ-0151-A-A-2      T14         20151008          7   41 106
121 16SAMP-12089 LIB-NDJ-0151-B-A-2      T14         20151008          7   41 106
122 16SAMP-12090 LIB-NDJ-0151-C-A-2      T14         20151008          7   41 106
115   16SAMP-9873    LIB-NDJ-0182-C-1      T17         20160114         12    2 204
123 16SAMP-12091 LIB-NDJ-0182-A-A-2      T17         20160114         12    2 204
124 16SAMP-12092 LIB-NDJ-0182-B-A-2      T17         20160114         12    2 204
125 16SAMP-12093 LIB-NDJ-0182-C-A-2      T17         20160114         12    2 204
```

Fjernelse af de resterende prøver fra den ene sekvenseringsomgang.

```
meta_fortyndet=meta_fortyndet[-which(meta_fortyndet$Seq_ID=="16SAMP-9873"),]
meta_fortyndet=meta_fortyndet[-which(meta_fortyndet$Seq_ID=="16SAMP-9778"),]
otu_bak_fortyndet=otu_bak_fortyndet[,-which(colnames(otu_bak_fortyndet)=="16SAMP-9778")]
otu_bak_fortyndet=otu_bak_fortyndet[,-which(colnames(otu_bak_fortyndet)=="16SAMP-9873")]
which(rowSums(otu_bak_fortyndet[,2:160])==0)
```

```
named integer(0)
```

Beregning af forskellen mellem største antal reads og mindste antal reads for prøverne inden for en triplikat.

```
sumreads_fortyndet=vector()
for(x in 1:length(meta_fortyndet$Seq_ID)){
sumreadsen=colSums(otu_bak_fortyndet[,2:160])[meta_fortyndet$Seq_ID[x]]
sumreads_fortyndet=c(sumreads_fortyndet,sumreadsen)
  }
meta_tr_fortyndet=meta_fortyndet
meta_tr_fortyndet$sum_reads=sumreads_fortyndet
diff_T14=vector()
diff_T16=vector()
diff_T17=vector()
max_T14=vector()
max_T16=vector()
max_T17=vector()
```

```
var_T14=vector()
var_T16=vector()
var_T17=vector()
for(x in 1:14){
tmp1=intersect(which(meta_tr_fortyndet$Reactor=="T14"),which(meta_tr_fortyndet$Data_Point==x))
tmp2=intersect(which(meta_tr_fortyndet$Reactor=="T16"),which(meta_tr_fortyndet$Data_Point==x))
tmp3=intersect(which(meta_tr_fortyndet$Reactor=="T17"),which(meta_tr_fortyndet$Data_Point==x))
diff14=diff(range(meta_tr_fortyndet$sum_reads[tmp1]))
diff_T14=c(diff_T14,diff14)
diff16=diff(range(meta_tr_fortyndet$sum_reads[tmp2]))
diff_T16=c(diff_T16,diff16)
diff17=diff(range(meta_tr_fortyndet$sum_reads[tmp3]))
diff_T17=c(diff_T17,diff17)
maxT14=which(meta_tr_fortyndet$sum_reads==max(meta_tr_fortyndet$sum_reads[tmp1]))
max_T14=c(max_T14,maxT14)
maxT16=which(meta_tr_fortyndet$sum_reads==max(meta_tr_fortyndet$sum_reads[tmp2]))
max_T16=c(max_T16,maxT16)
maxT17=which(meta_tr_fortyndet$sum_reads==max(meta_tr_fortyndet$sum_reads[tmp3]))
max_T17=c(max_T17,maxT17)
var14=var(meta_tr_fortyndet$sum_reads[tmp1])
var_T14=c(var_T14,var14)
var16=var(meta_tr_fortyndet$sum_reads[tmp3])
var_T16=c(var_T16,var16)
var17=var(meta_tr_fortyndet$sum_reads[tmp3])
var_T17=c(var_T17,var17)
}
c(max(diff_T14),max(diff_T16),max(diff_T17))
Triplikat_max_diff_sek=cbind(Datapoint=c(1:14),T14=diff_T14,T16=diff_T16,T17=diff_T17)
Triplikat_max_diff_sek
```

```
 Datapoint   T14   T16   T17
 [1,]         1 12271  2801  8492
 [2,]         2 16242  1184  7299
 [3,]         3 17345  4898  3953
 [4,]         4 24910 14529  9951
 [5,]         5 20255  7869  5665
 [6,]         6   913 17461  8329
 [7,]         7 22094 13596  2963
 [8,]         8  3089  7930  7269
 [9,]         9  9535 11340  3515
[10,]        10 15554 14376  5178
[11,]        11 10753  3788 14647
[12,]        12  5338 15057 15862
[13,]        13 10527 13040  9058
[14,]        14 10465 24265 21134
```

```
var_T14;var_T16;var_T17
Triplikat_var=cbind(Datapoint=c(1:14),T14=var_T14,T16=var_T16,T17=var_T17)
Triplikat_var
```

```
     Datapoint          T14       T16       T17
 [1,]         1   45147787.0  19678724  19678724
 [2,]         2   74839606.3  13318854  13318854
 [3,]         3   75305713.0   4001782   4001782
 [4,]         4  190662233.3  32934654  32934654
 [5,]         5  132633758.3   8046553   8046553
 [6,]         6     416784.5  22419562  22419562
 [7,]         7  140482449.3   2210611   2210611
 [8,]         8    4770960.5  13253634  13253634
 [9,]         9   22754173.0   3219640   3219640
[10,]        10   65938434.3   7486284   7486284
[11,]        11   30054379.0  57926096  57926096
[12,]        12    7231861.0  80537786  80537786
[13,]        13   31089026.3  14988589  14988589
[14,]        14   27454423.0 139605601 139605601
```

Bestemmelse af forskellen mellem mindste og største antal reads for triplikatet med størst forskel pr. reaktor.

```
c(max(diff_T14),max(diff_T16),max(diff_T17))
```

```
[1] 24910 24265 21134
```

Konstruktion af dataframe med en prøve fra hver triplikat.

```
sample_fortyndet=c(max_T14,max_T16,max_T17,which(as.numeric(meta_tr_fortyndet$Data_Point)>14))
meta_tr_fortyndet=meta_tr_fortyndet[sample_fortyndet,]
meta_tr_fortyndet=meta_tr_fortyndet[order(meta_tr_fortyndet$Re,meta_tr_fortyndet$Day),]
meta_fortyndet=meta_fortyndet[sample_fortyndet,]
meta_fortyndet=meta_fortyndet[order(meta_fortyndet$Re,meta_fortyndet$Day),]
```

```
save(meta_fortyndet,file="meta_fortyndet.Rda")
tmp01=c(colnames(otu_bak_fortyndet)[1],intersect(meta_fortyndet$Seq_ID,colnames(otu_bak_
    fortyndet)),colnames(otu_bak_fortyndet)[161:167])
otu_bak_fortyndet=otu_bak_fortyndet[,tmp01] # indeholder 3343 OTU'er
otu_bak_fortyndetantalotu=otu_bak_fortyndet[apply(otu_bak[,2:77], 1, max)>0,] #indeholder af
    3288  OTU'er
save(otu_bak_fortyndet,file="otu_bak_fortyndet.Rda" )
```

Udvælgelse af OTU'er til modellering.

```
otu_bak_kfortyndet=otu_bak_fortyndet
for(x in 2:77){
  otu_bak_kfortyndet[x]=sqrt(otu_bak_kfortyndet[x])
}
reaktor_sample <- split(meta_fortyndet$Seq_ID,meta_fortyndet$Reactor)
OTU_reaktor <- list()
reactor_names <- c("T14","T16","T17")
  for(x in reactor_names){ OTU_reaktor[[x]]$abundance <-   apply(otu_bak_kfortyndet[,reaktor_
      sample[[x]]],2,function(x) 1000*x/sum(x))
  a=OTU_reaktor[[x]]$median_overEn=apply(OTU_reaktor[[x]]$ab,1,median)>1
  b=OTU_reaktor[[x]]$median_overFem=apply(OTU_reaktor[[x]]$ab,1,median)>5
  c=OTU_reaktor[[x]]$median_overTi=apply(OTU_reaktor[[x]]$ab,1,median)>10
  OTU_reaktor[[x]]$total_median=c(sum(a),sum(b),sum(c))
  d=OTU_reaktor[[x]]$max_overEn=apply(OTU_reaktor[[x]]$ab,1,max)>1
  e=OTU_reaktor[[x]]$max_overFem=apply(OTU_reaktor[[x]]$ab,1,max)>5
  f=OTU_reaktor[[x]]$max_overTi=apply(OTU_reaktor[[x]]$ab,1,max)>10
  OTU_reaktor[[x]]$total_max=c(sum(d),sum(e),sum(f))
  g=OTU_reaktor[[x]]$mean_overEn=apply(OTU_reaktor[[x]]$ab,1,mean)>1
  h=OTU_reaktor[[x]]$mean_overFem=apply(OTU_reaktor[[x]]$ab,1,mean)>5
  i=OTU_reaktor[[x]]$mean_overTi=apply(OTU_reaktor[[x]]$ab,1,mean)>10
  OTU_reaktor[[x]]$total_mean=c(sum(g),sum(h),sum(i))
  }
max_union1=OTU_reaktor[[1]]$max_overEn
max_union5=OTU_reaktor[[1]]$max_overFem
max_union10=OTU_reaktor[[1]]$max_overTi
max_intersect1=OTU_reaktor[[1]]$max_overEn
max_intersect5=OTU_reaktor[[1]]$max_overFem
max_intersect10=OTU_reaktor[[1]]$max_overTi
median_union1=OTU_reaktor[[1]]$median_overEn
median_union5=OTU_reaktor[[1]]$median_overFem
median_union10=OTU_reaktor[[1]]$median_overTi
median_intersect1=OTU_reaktor[[1]]$median_overEn
median_intersect5=OTU_reaktor[[1]]$median_overFem
median_intersect10=OTU_reaktor[[1]]$median_overTi
mean_union1=OTU_reaktor[[1]]$mean_overEn
mean_union5=OTU_reaktor[[1]]$mean_overFem
mean_union10=OTU_reaktor[[1]]$mean_overTi
mean_intersect1=OTU_reaktor[[1]]$mean_overEn
mean_intersect5=OTU_reaktor[[1]]$mean_overFem
mean_intersect10=OTU_reaktor[[1]]$mean_overTi
for(i in 2:3){max_union1= max_union1|OTU_reaktor[[i]]$max_overEn
              max_union5=max_union5|OTU_reaktor[[i]]$max_overFem
              max_union10=max_union10|OTU_reaktor[[i]]$max_overTi
              max_intersect1=max_intersect1&OTU_reaktor[[i]]$max_overEn
              max_intersect5=max_intersect5&OTU_reaktor[[i]]$max_overFem
              max_intersect10=max_intersect10&OTU_reaktor[[i]]$max_overTi
              median_union1= median_union1|OTU_reaktor[[i]]$median_overEn
              median_union5=median_union5|OTU_reaktor[[i]]$median_overFem
              median_union10=median_union10|OTU_reaktor[[i]]$median_overTi
              median_intersect1= median_intersect1&OTU_reaktor[[i]]$median_overEn
              median_intersect5=median_intersect5&OTU_reaktor[[i]]$median_overFem
              median_intersect10=median_intersect10&OTU_reaktor[[i]]$median_overTi
              mean_union1= mean_union1|OTU_reaktor[[i]]$mean_overEn
              mean_union5=mean_union5|OTU_reaktor[[i]]$mean_overFem
              mean_union10=mean_union10|OTU_reaktor[[i]]$mean_overTi
              mean_intersect1= mean_intersect1&OTU_reaktor[[i]]$mean_overEn
              mean_intersect5=mean_intersect5&OTU_reaktor[[i]]$mean_overFem
              mean_intersect10=mean_intersect10&OTU_reaktor[[i]]$mean_overTi
              }
union_max=c(sum(max_union1),sum(max_union5),sum(max_union10))
intersect_max=c(sum(max_intersect1),sum(max_intersect5),sum(max_intersect10))
union_median=c(sum(median_union1),sum(median_union5),sum(median_union10))
intersect_median=c(sum(median_intersect1),sum(median_intersect5),sum(median_intersect10))
union_mean=c(sum(mean_union1),sum(mean_union5),sum(mean_union10))
intersect_mean=c(sum(mean_intersect1),sum(mean_intersect5),sum(mean_intersect10))
sumreads=vector()
otu_tmp=otu_bak_fortyndet[,2:77]
sumreads=vector()
for(x in 1:length(otu_bak_fortyndet$`16SAMP-16995`)){
sumreadsen=rowSums(otu_tmp)[x]
sumreads=c(sumreads,sumreadsen)
}
otu_rowsum=otu_bak_fortyndet
otu_rowsum$sum_reads=sumreads
```

```
procentmax_union=c(sum(otu_rowsum$sum_reads[max_union1]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[max_union5]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[max_union10]/sum(otu_rowsum$sum_reads)))
procentmax_intersect=c(sum(otu_rowsum$sum_reads[max_intersect1]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[max_intersect5]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[max_intersect10]/sum(otu_rowsum$sum_reads)))
procentmedian_union=c(sum(otu_rowsum$sum_reads[median_union1]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[median_union5]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[median_union10]/sum(otu_rowsum$sum_reads)))
procentmedian_intersect=c(sum(otu_rowsum$sum_reads[median_intersect1]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[median_intersect5]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[median_intersect10]/sum(otu_rowsum$sum_reads)))
procentmean_union=c(sum(otu_rowsum$sum_reads[mean_union1]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[mean_union5]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[mean_union10]/sum(otu_rowsum$sum_reads)))
procentmean_intersect=c(sum(otu_rowsum$sum_reads[mean_intersect1]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[mean_intersect5]/sum(otu_rowsum$sum_reads)),sum(otu_rowsum$sum_reads[mean_intersect10]/sum(otu_rowsum$sum_reads)))
rensning_max=rbind("T14"=OTU_reaktor[[1]]$total_max,"T16"=OTU_reaktor[[2]]$total_max,"T17"=OTU_reaktor[[3]]$total_max, union_max,  procentmax_union, intersect_max, procentmax_intersect)
rensning_median=rbind("T14"=OTU_reaktor[[1]]$total_median,"T16"=OTU_reaktor[[2]]$total_median,"T17"=OTU_reaktor[[3]]$total_median, union_median,  procentmedian_union, intersect_median, procentmedian_intersect)
rensning_mean=rbind("T14"=OTU_reaktor[[1]]$total_mean,"T16"=OTU_reaktor[[2]]$total_mean,"T17"=OTU_reaktor[[3]]$total_mean, union_mean,  procentmean_union, intersect_mean, procentmean_intersect)
rensning_max
```

```
                          [,1]        [,2]        [,3]
T14                561.0000000 47.0000000 16.0000000
T16                788.0000000 49.0000000 21.0000000
T17                569.0000000 44.0000000 13.0000000
union_max          885.0000000 58.0000000 22.0000000
procentmax_union     0.9768556  0.7922690  0.6694100
intersect_max      451.0000000 38.0000000 13.0000000
procentmax_intersect   0.9544235  0.7525722  0.5951976
```

```
rensning_median
```

```
                            [,1]        [,2]      [,3]
T14                  193.0000000 16.0000000 5.0000000
T16                  197.0000000 18.0000000 5.0000000
T17                  207.0000000 19.0000000 6.0000000
union_median         231.0000000 20.0000000 6.0000000
procentmedian_union    0.9128424  0.6724402 0.4927791
intersect_median     169.0000000 14.0000000 5.0000000
procentmedian_intersect   0.8861080  0.6146220 0.4650543
```

```
rensning_mean
```

```
                          [,1]        [,2]      [,3]
T14                214.0000000 19.0000000 5.0000000
T16                201.0000000 19.0000000 7.0000000
T17                209.0000000 21.0000000 6.0000000
union_mean         238.0000000 22.0000000 7.0000000
procentmean_union    0.9190563  0.6878834 0.5139972
intersect_mean     172.0000000 18.0000000 5.0000000
procentmean_intersect   0.8943984  0.6585425 0.4650543
```

Konstruktion af nyt dataframe.

```
otu_bak_final=otu_bak_fortyndet[median_intersect5,]
otu_bak_kfinal=otu_bak_kfortyndet[median_intersect5,]
save(otu_bak_final, file = "otu_bak_final.Rda")
save(otu_bak_kfinal, file = "otu_bak_kfinal.Rda")
```

A.4 PCA, heatmap og spline

For at kører kodning skal pakkerne ”ggplot2”, og ”ampvis2”hentes. Datasættene `otu_bak` og `meta` er de redigerede datasæt med ændringerne som ses ved A.2.

```
library(ggplot2)
library(ampvis2)
data_bakterier <- amp_load(otutable = otu_bak,metadata = meta)
rarefrac_bakterier=amp_rarecurve(data=data_bakterier)
plot(rarefrac_bakterier)
```

Kodning for PCA er følgende:

```
data_bak_fortyndet <- amp_load(otutable = otu_bak_fortyndet,metadata = meta_fortyndet)
pca=amp_ordinate(data_bak_fortyndet, sample_label_by = "Data_Point", sample_color_by = "Reactor
    ")+ylim(-0.4,0.4)+xlim(-0.4,0.4) +
  theme_bw() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
pca
```

Kodning for heatmap er følgende:

```
subdataT14=amp_subset_samples(data_bak_fortyndet,Reactor %in% "T14")
subdatatriT14=amp_subset_samples(subdataT14,Data_Point %in% c(1,2,3,4,5,6,7,8,9,10,11,12,13,14)
    )
subdatatriT14_1=amp_subset_samples(subdataT14,Data_Point %in% c(1,2,3,4,5,6,7))
subdatatriT14_2=amp_subset_samples(subdataT14,Data_Point %in% c(8,9,10,11,12,13,14))
T14_1=amp_heatmap(subdatatriT14_1,
                        tax_aggregate="Species",
            group=c("Data_Point", "Seq_ID"),
                        tax_show=20,
                        tax_empty = "best",
                        color_vector = c("White", "red"),
                        plot_colorscale = "sqrt")
T14_1
T14_2=amp_heatmap(subdatatriT14_2,
                        tax_aggregate="Species",
            group=c("Data_Point", "Seq_ID"),
                        tax_show=20,
                        tax_empty = "best",
                        color_vector = c("White", "red"),
                        plot_colorscale = "sqrt")
T14_2
subdataT16=amp_subset_samples(data_bak_fortyndet,Reactor %in% "T16")
subdatatriT16=amp_subset_samples(subdataT16,Data_Point %in% c(1,2,3,4,5,6,7,8,9,10,11,12,13,14)
    )
subdatatriT16_1=amp_subset_samples(subdataT16,Data_Point %in% c(1,2,3,4,5,6,7))
subdatatriT16_2=amp_subset_samples(subdataT16,Data_Point %in% c(8,9,10,11,12,13,14))
T16_1=amp_heatmap(subdatatriT16_1,
                        tax_aggregate="Species",
            group=c("Data_Point", "Seq_ID"),
                        tax_show=20,
                        tax_empty = "best",
                        color_vector = c("White", "red"),
                        plot_colorscale = "sqrt")
T16_1
T16_2=amp_heatmap(subdatatriT16_2,
                        tax_aggregate="Species",
            group=c("Data_Point", "Seq_ID"),
                        tax_show=20,
                        tax_empty = "best",
                        color_vector = c("White", "red"),
                        plot_colorscale = "sqrt")
T16_2
subdataT17=amp_subset_samples(data_bak_fortyndet,Reactor %in% "T17")
subdatatriT17=amp_subset_samples(subdataT17,Data_Point %in% c(1,2,3,4,5,6,7,8,9,10,11,12,13,14)
    )
subdatatriT17_1=amp_subset_samples(subdataT17,Data_Point %in% c(1,2,3,4,5,6,7))
subdatatriT17_2=amp_subset_samples(subdataT17,Data_Point %in% c(8,9,10,11,12,13,14))
T17_1=amp_heatmap(subdatatriT17_1,
                        tax_aggregate="Species",
            group=c("Data_Point", "Seq_ID"),
                        tax_show=20,
                        tax_empty = "best",
                        color_vector = c("White", "red"),
                        plot_colorscale = "sqrt")
T17_1
T17_2=amp_heatmap(subdatatriT17_2,
                        tax_aggregate="Species",
            group=c("Data_Point", "Seq_ID"),
                        tax_show=20,
                        tax_empty = "best",
                        color_vector = c("White", "red"),
                        plot_colorscale = "sqrt")
T17_2
```

Til spline er datasættet konstrueret i A.3 anvendt:

```
load("otu_bak_kfinal.Rda")
otu_bak_kfinal_n=apply(otu_bak_kfinal[,2:77], 2, function(x) 1000*x/sum(x))
load("meta_fortyndet.Rda")
library(splines)
library(ggplot2)
library(gridExtra)
library(grid)
```

Kodning af spline for OTU'er i reaktor T14.

```
t=meta_fortyndet$Day[1:27]
#For at placere 15 knuder ligeligt mellem mindste observation og stoerste observation er kno <-
     seq(t[1],t[27],length=15) anvendt. Saettes knuderne til at vaere lig med observationerne
    fratrukket den mindste og oeverste observation anvendes kno <- t[3:25].
kno <-  quantile(t,c((1:15)/16))
t0 <- seq(t[1],t[27],length=27)
t1 <- seq(t[1],t[27],length=1000)
otu_T14=otu_bak_kfinal_n[,meta_fortyndet$Seq_ID[which(meta_fortyndet$Reactor=="T14")]]

fitmodel_list=list()
y0hat_list=list()
yhat_list=list()
#For at genere en fjerde grads spline er deg=4 anvendt
for(i in 1:14){
  fitmodel_list[[i]]=lm(as.numeric(otu_T14[i,])~ bs(t,kno=kno,deg=3,Bound=t[c(1,27)]))
y0hat_list[[i]] <- predict(fitmodel_list[[i]],data.frame(t=t0))
yhat_list[[i]] <- predict(fitmodel_list[[i]],data.frame(t=t1))
}

###Plot splines vdh ggplot#############################################################
T14_y0hat=list()
for(i in 1:14){T14_y0hat[[i]]=y0hat_list[[i]]
               T14_y0hat[[i+14]]=as.numeric(otu_T14[i,])
}
               T14_y0hat[[29]]=t
               T14_y0hat[[30]]=t0
T14_yhat=list()
for(i in 1:14){T14_yhat[[i]]=yhat_list[[i]]}
                 T14_yhat[[15]]=t1

#konverterer det til en dataframe
T14_y0hat=data.frame((sapply(T14_y0hat,c)))
T14_yhat=data.frame((sapply(T14_yhat,c)))

########Dannelse af otu dataframe#############
OTU_splineT14=data.frame(t(sapply(y0hat_list,c)))
for(i in 1:27){names(OTU_splineT14)[i]=paste0("T14_",i)}

OTU_splineT14=cbind(otu_bak_kfinal[1],OTU_splineT14)

#############Plot af spline vdh ggplot#########
plot_list=list()
for(i in 1:14){
  plot_list[[i]] <-
    ggplot() +
    geom_point(data = T14_y0hat, aes_string(x="X30", y=paste0("X",i))) +
    geom_point(data = T14_y0hat, aes_string(x="X29", y=paste0("X", i+14)), colour="red")+
geom_line(data=T14_yhat,aes_string(x="X15",y=paste0("X",i)))+
    xlab("Dag") +ylab("Transformeret antal") +ggtitle(otu_bak_kfinal[i,1])
  }

SplineT14_1=grid.arrange(
  plot_list[[7]], plot_list[[6]],plot_list[[4]], plot_list[[9]], plot_list[[3]],plot_list[[5]],
       plot_list[[8]], plot_list[[14]]
             , ncol=2)

SplineT14_2=grid.arrange(
plot_list[[10]], plot_list[[13]], plot_list[[1]], plot_list[[12]], plot_list[[11]],plot_list
     [[2]],
              ncol=2)
```

Kodning til spline for reaktor T16.

```
t16=meta_fortyndet$Day[28:50]
#For at placere knuderne ved percentilerne saettes kno16 <-quantile(t16,c((1:13)/14))
kno16 <- seq(t[1],t[23],length=13)
otu_T16=otu_bak_kfinal_n[,meta_fortyndet$Seq_ID[which(meta_fortyndet$Reactor=="T16")]]

fitmodel_listT16=list()
y0hat_listT16=list()
yhat_listT16=list()
for(i in 1:14){
  fitmodel_listT16[[i]]=lm( as.numeric(otu_T16[i,])~ bs(t16,kno=kno16,deg=3,sBound=t16[c(1,23)
      ]))
y0hat_listT16[[i]] <- predict(fitmodel_listT16[[i]],data.frame(t16=t0))
yhat_listT16[[i]] <- predict(fitmodel_listT16[[i]],data.frame(t16=t1))
}

T16_yhat=list()
T16_y0hat_1=list()
T16_y0hat=list()
for(i in 1:14){T16_yhat[[i]]=yhat_listT16[[i]]
               }
```

```
                T16_yhat[[15]]=t1
for(i in 1:14){T16_y0hat_1[[i]]=y0hat_listT16[[i]]
                  }
               T16_y0hat_1[[15]]=t0

for(i in 1:14){T16_y0hat[[i]]=as.numeric(otu_T16[i,])
               }
               T16_y0hat[[15]]=t16

T16_yhat=data.frame((sapply(T16_yhat,c)))
T16_y0hat_1=data.frame((sapply(T16_y0hat_1,c)))
T16_y0hat=data.frame((sapply(T16_y0hat,c)))
###Dannelse af dataframe for OTU spline######################
OTU_splineT16=data.frame(t(sapply(y0hat_listT16,c)))
for(i in 1:27){names(OTU_splineT16)[i]=paste0("T16_",i)}
##Plot ###########################################################################
plot_listT16=list()
for(i in 1:14){
  plot_listT16[[i]] <-
    ggplot() +
    geom_point(data = T16_y0hat_1, aes_string(x="X15", y=paste0("X",i))) +
    geom_point(data = T16_y0hat, aes_string(x="X15", y=paste0("X", i)), colour="red")+
geom_line(data=T16_yhat,aes_string(x="X15",y=paste0("X",i)))+
    xlab("Dag") +ylab("Transformeret antal") +ggtitle(otu_bak_kfinal[i,1])
  }

SplineT16_1=grid.arrange(
  plot_listT16[[7]], plot_listT16[[6]],plot_listT16[[4]], plot_listT16[[9]], plot_listT16[[3]],
      plot_listT16[[5]], plot_listT16[[11]],plot_listT16[[2]]
             , ncol=2)

SplineT16_2=grid.arrange(
  plot_listT16[[8]], plot_listT16[[14]], plot_listT16[[10]], plot_listT16[[13]], plot_listT16
      [[1]], plot_listT16[[12]],
             ncol=2)
```

Kodning af spline for OTU'er i reaktor T17.

```
t17=meta_fortyndet$Day[51:76]
#For at placerer knuderne ved percentilerne saettes kno17 <-quantile(t17,c((1:14)/15))
kno17 <- seq(t[1],t[26],length=14)

otu_T17=otu_bak_kfinal_n[,meta_fortyndet$Seq_ID[which(meta_fortyndet$Reactor=="T17")]]

fitmodel_listT17=list()
y0hat_listT17=list()
yhat_listT17=list()
for(i in 1:14){
  fitmodel_listT17[[i]]=lm( as.numeric(otu_T17[i,])~ bs(t17,kno=kno17,deg=3,Bound=t17[c(1,26)])
      )
y0hat_listT17[[i]] <- predict(fitmodel_listT17[[i]],data.frame(t17=t0))
yhat_listT17[[i]] <- predict(fitmodel_listT17[[i]],data.frame(t17=t1))
}

T17_y0hat_1=list()
T17_y0hat=list()
T17_yhat=list()
for(i in 1:14){T17_y0hat_1[[i]]=y0hat_listT17[[i]]
                  }
               T17_y0hat_1[[15]]=t0

for(i in 1:14){T17_y0hat[[i]]=as.numeric(otu_T17[i,])
               }
               T17_y0hat[[15]]=t17

for(i in 1:14){T17_yhat[[i]]=yhat_listT17[[i]]}
                 T17_yhat[[15]]=t1

T17_y0hat_1=data.frame((sapply(T17_y0hat_1,c)))
T17_y0hat=data.frame((sapply(T17_y0hat,c)))
T17_yhat=data.frame((sapply(T17_yhat,c)))

###Dannelse af dataframe for OTU spline######################
OTU_splineT17=data.frame(t(sapply(y0hat_listT17,c)))
for(i in 1:27){names(OTU_splineT17)[i]=paste0("T17_",i)}

##Plot ###########################################################################
plot_listT17=list()
for(i in 1:14){
  plot_listT17[[i]] <-
    ggplot() +
    geom_point(data = T17_y0hat_1, aes_string(x="X15", y=paste0("X",i))) +
    geom_point(data = T17_y0hat, aes_string(x="X15", y=paste0("X", i)), colour="red")+
geom_line(data=T17_yhat,aes_string(x="X15",y=paste0("X",i)))+
    xlab("Dag") +ylab("Transformeret antal") +ggtitle(otu_bak_kfinal[i,1])
```

```
47   }
48
49 SplineT17_1=grid.arrange(
50   plot_listT17[[7]], plot_listT17[[6]],plot_listT17[[4]], plot_listT17[[9]], plot_listT17[[3]],
       plot_listT17[[5]], plot_listT17[[11]],plot_listT17[[2]]
51              , ncol=2)
52
53 SplineT17_2=grid.arrange(
54   plot_listT17[[8]], plot_listT17[[14]], plot_listT17[[10]], plot_listT17[[13]], plot_listT17
       [[1]], plot_listT17[[12]],
55              ncol=2)
```

Dannelse af dataframe med ækvidistante tidspunkter.

```
1 OTU_spline=cbind(OTU_splineT14,OTU_splineT16,OTU_splineT17, otu_bak_kfinal[78:84])
2 Akvivalent_tider=as.data.frame(t0)
3 names(Akvivalent_tider)[1]="Day"
4 save(OTU_spline,file="OTU_spline.Rda")
5 save(Akvivalent_tider,file="Akvivalent_tider.Rda")
```

A.5 Modellering af Lotka Volterra

Datasættet som indlæses er dannet i appendiks F og A.4.

```
1 load("A_ny.Rda") # Er "overgangsmatricen" for VARX(1)
2 load("OTU_spline.Rda")
3 load("Intercept_A_ny.Rda") #Er intercept til VARX(1)
4 data_samlet=as.data.frame(t(OTU_spline))
5 data_samlet[] <- lapply(data_samlet, as.character)
6 colnames(data_samlet) <- data_samlet[1, ]
7 data_samlet <- data_samlet[-1 ,]
8 T14_data=data_samlet[1:27,]
9 T14_data_matrix=data.matrix(T14_data, rownames.force = NA)
```

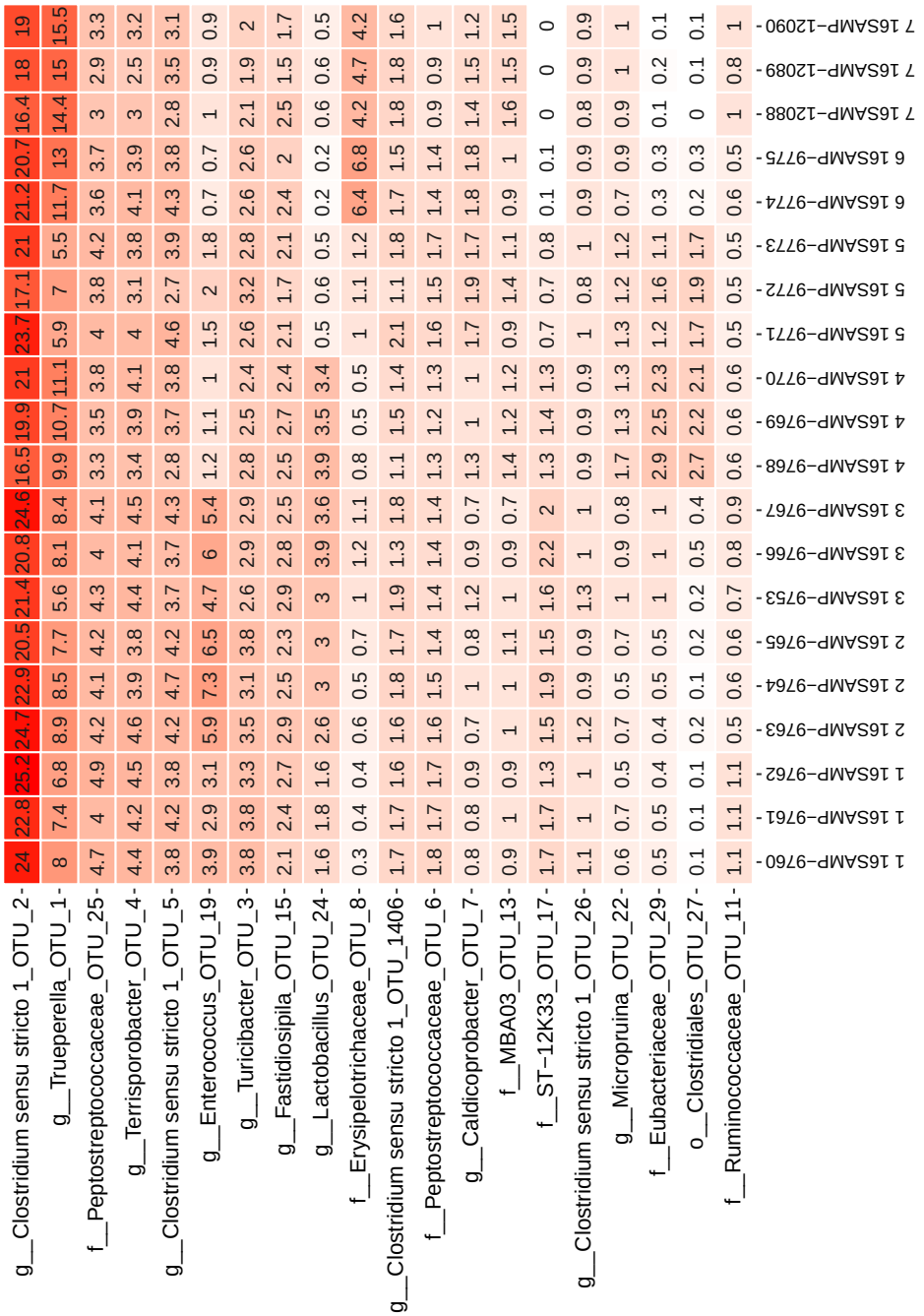
```
1 Intercept_A_ny=as.matrix(Intercept_A_ny)
2 rownames(Intercept_A_ny)=names(T14_data)
3 colnames(Intercept_A_ny)="Intercept"
```

```
1 B_ny=A_ny
2 diag(B_ny)=0
3 test=apply(B_ny,1,function(x) sum(0!=x)==0)
4 test1=test&apply(B_ny,2,function(x) sum(0!=x)==0)
5 arx=A_ny[!test1,!test1]
6 arx_intercept=as.matrix(Intercept_A_ny[rownames(Intercept_A_ny) %in% rownames(arx), ])
7 arx_mu=(apply(T14_data_matrix, 2, function(x) mean(x)))[rownames(arx)]
```

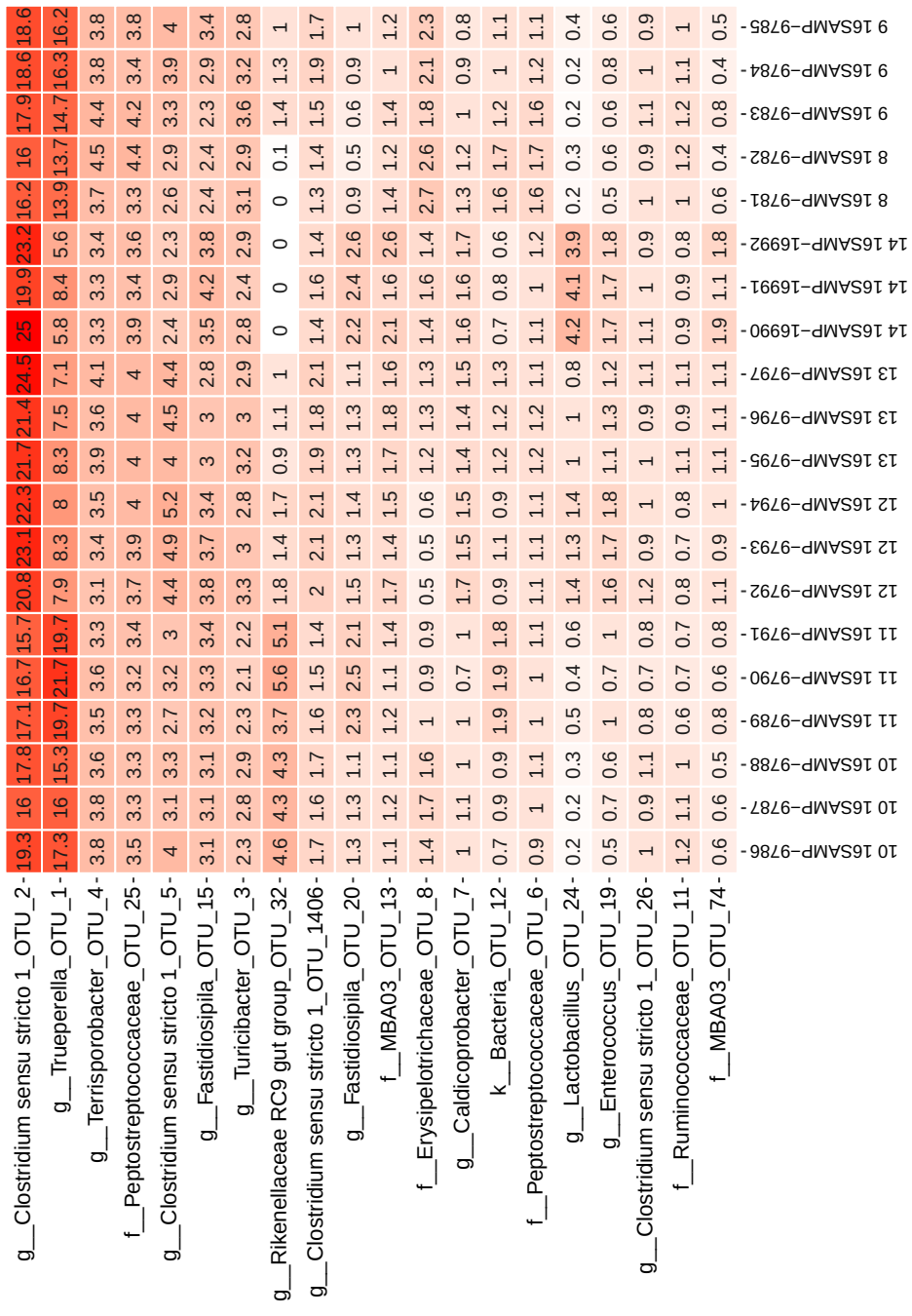
```
1 alfa=vector()
2 for(x in 1:length(arx_intercept)){
3 alfa[x]=arx_intercept[x]/arx_mu[x]
4 }
5 alfa=as.matrix(alfa)
6 row.names(alfa)=row.names(arx)
7 colnames(alfa)="alfa"
```

```
1 I <- matrix(0, 9, 9)
2 diag(I) <- 1
3 tmp=list()
4 for(x in 1:9){
5 tmp[[x]]=(arx[x,]-I[x,])*1/arx_mu[x]
6 }
7 Lotka_M=matrix(t(unlist(tmp)),ncol=9, byrow = TRUE)
8 rownames(Lotka_M)=row.names(arx)
9 colnames(Lotka_M)=colnames(arx)
```

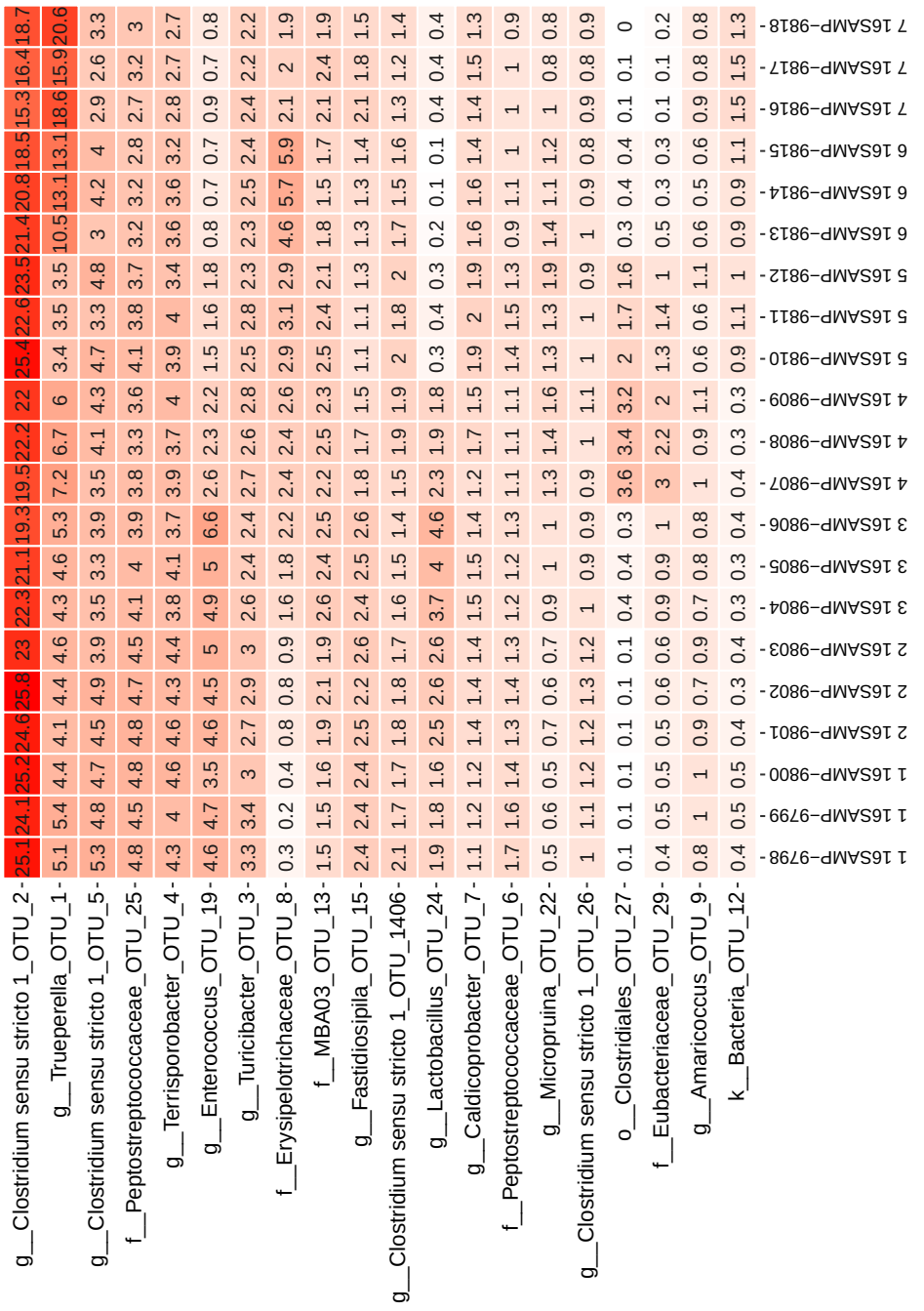
B Heatmaps



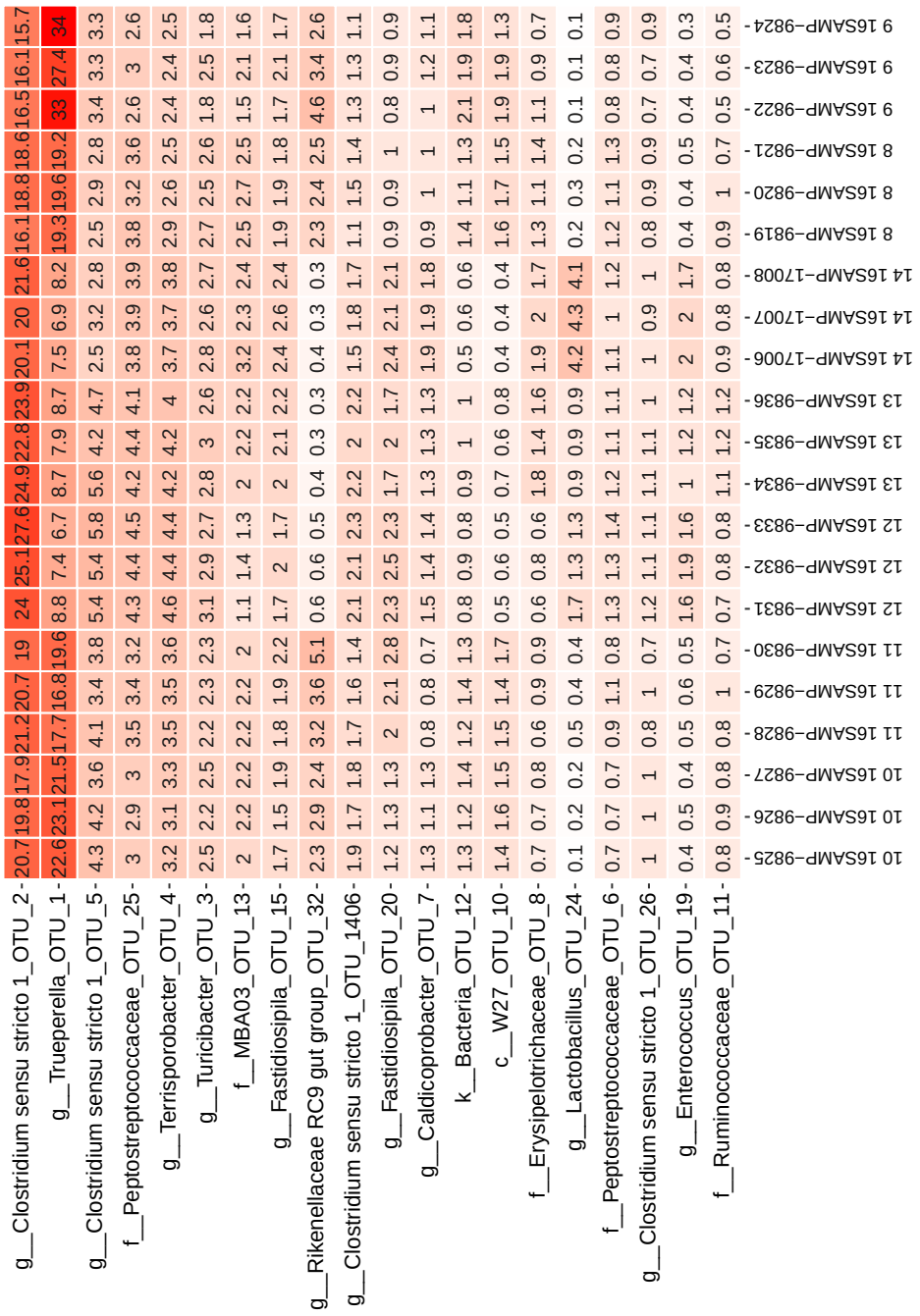
Figur 15: Heatmap over prøverne taget på prøvetagningsdag 1-7 i reaktor T14.



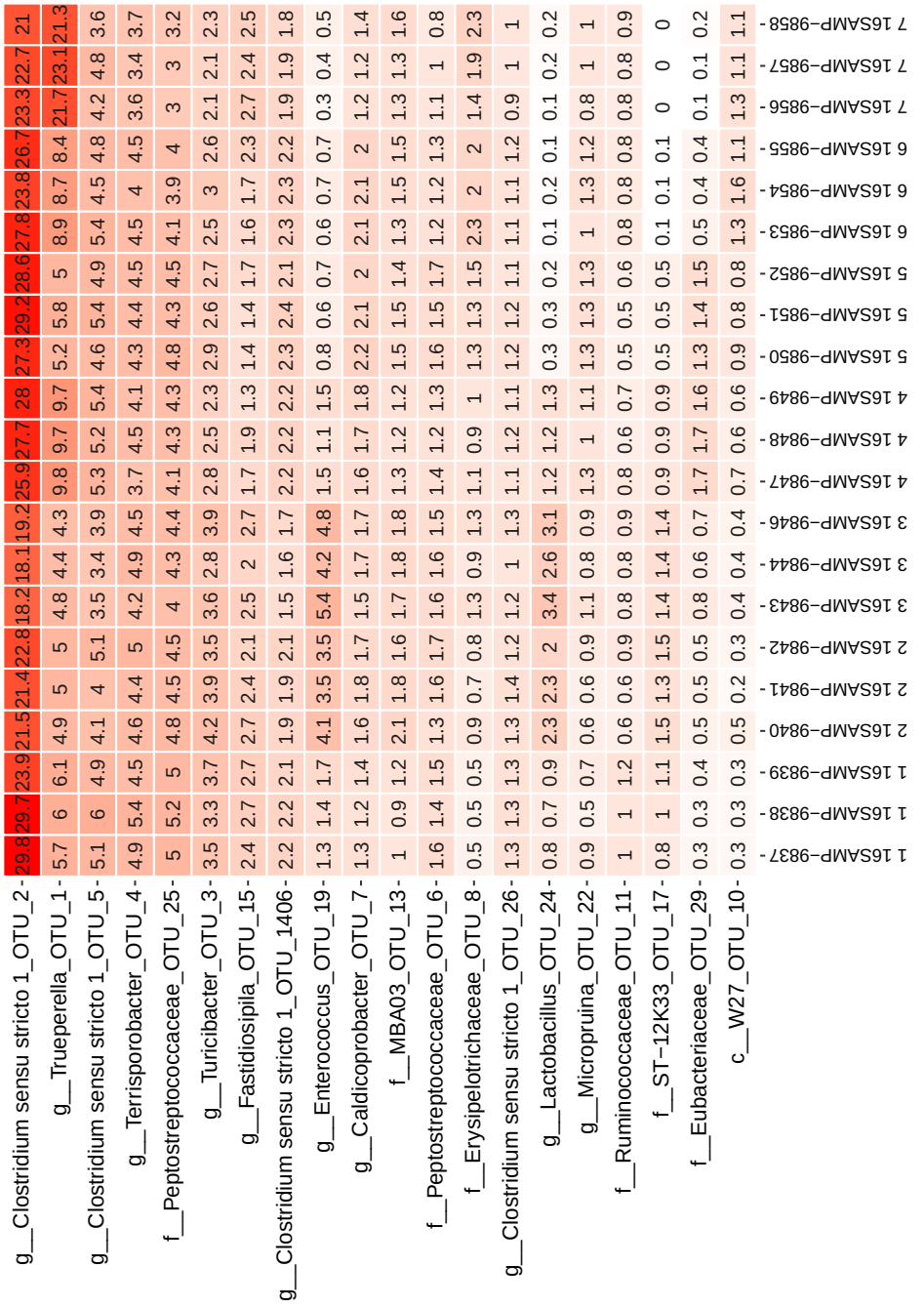
Figur 16: Heatmap over prøverne taget på prøvetagningsdag 8-14 i reaktor T14.



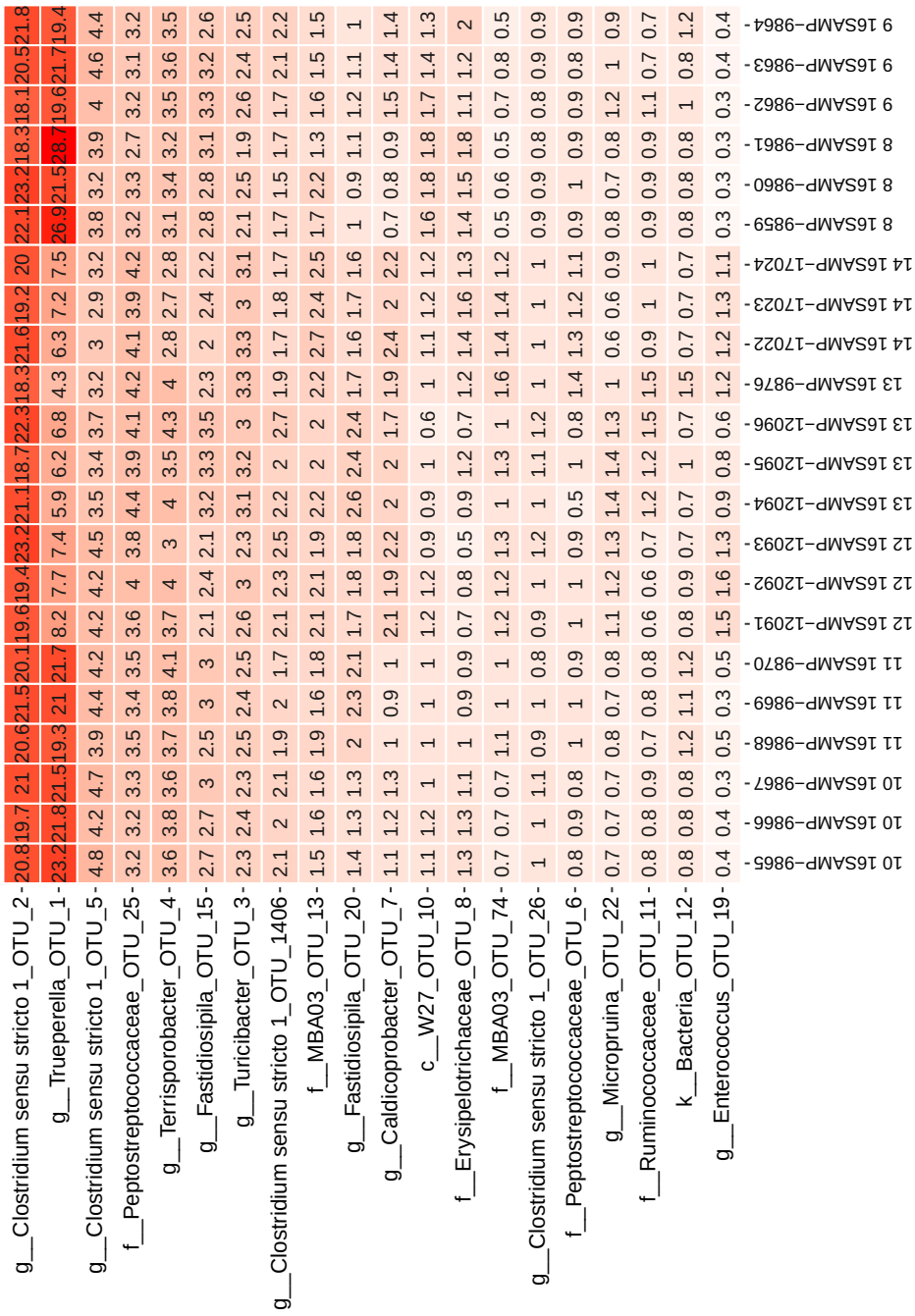
Figur 17: Heatmap over prøverne taget på prøvetagningsdag 1-7 i reaktor T16.



Figur 18: Heatmap over prøverne taget på prøvetagningsdag 8-14 i reaktor T16.

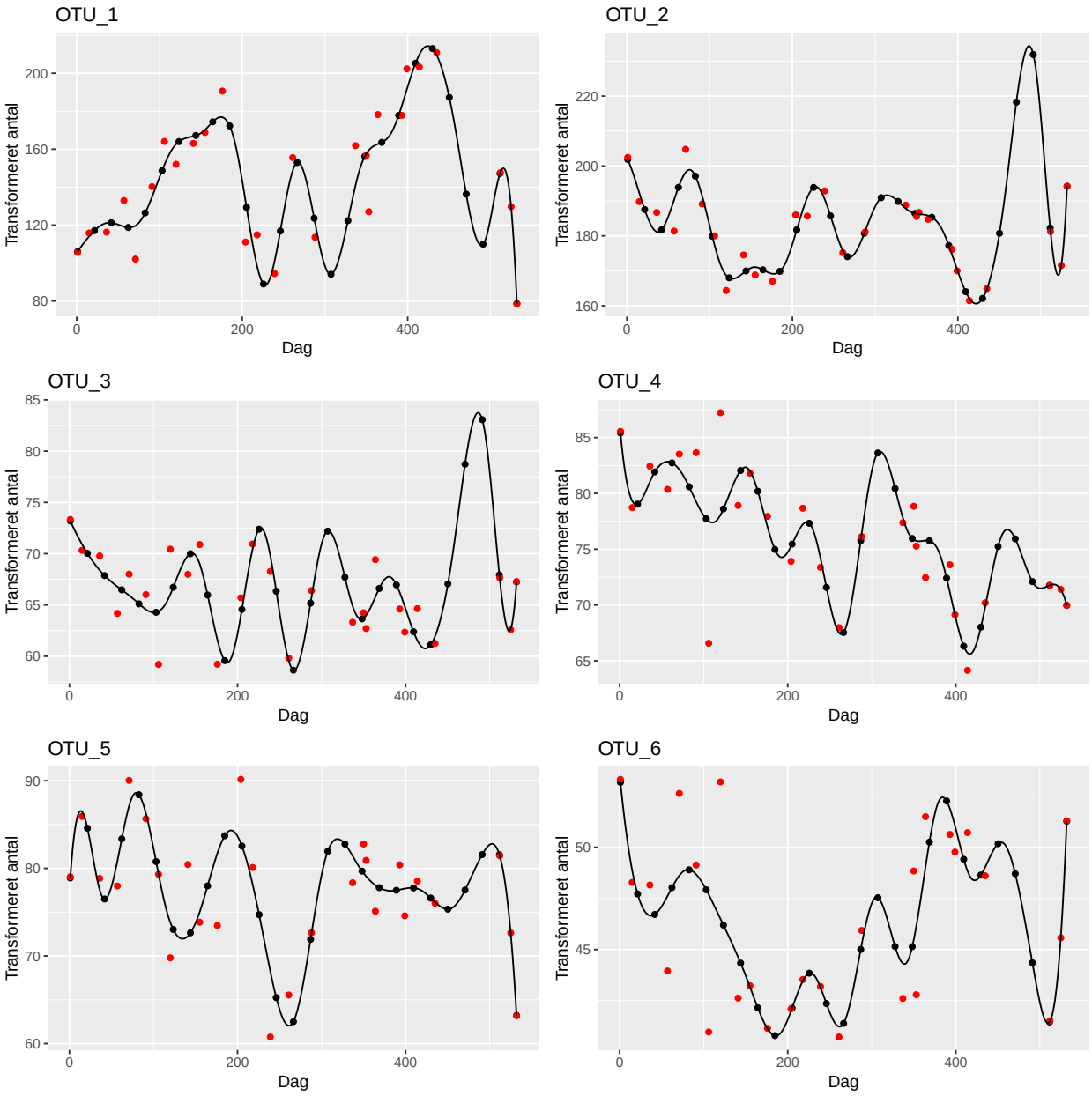


Figur 19: Heatmap over prøverne taget på prøvetagningsdag 1-7 i reaktor T17.

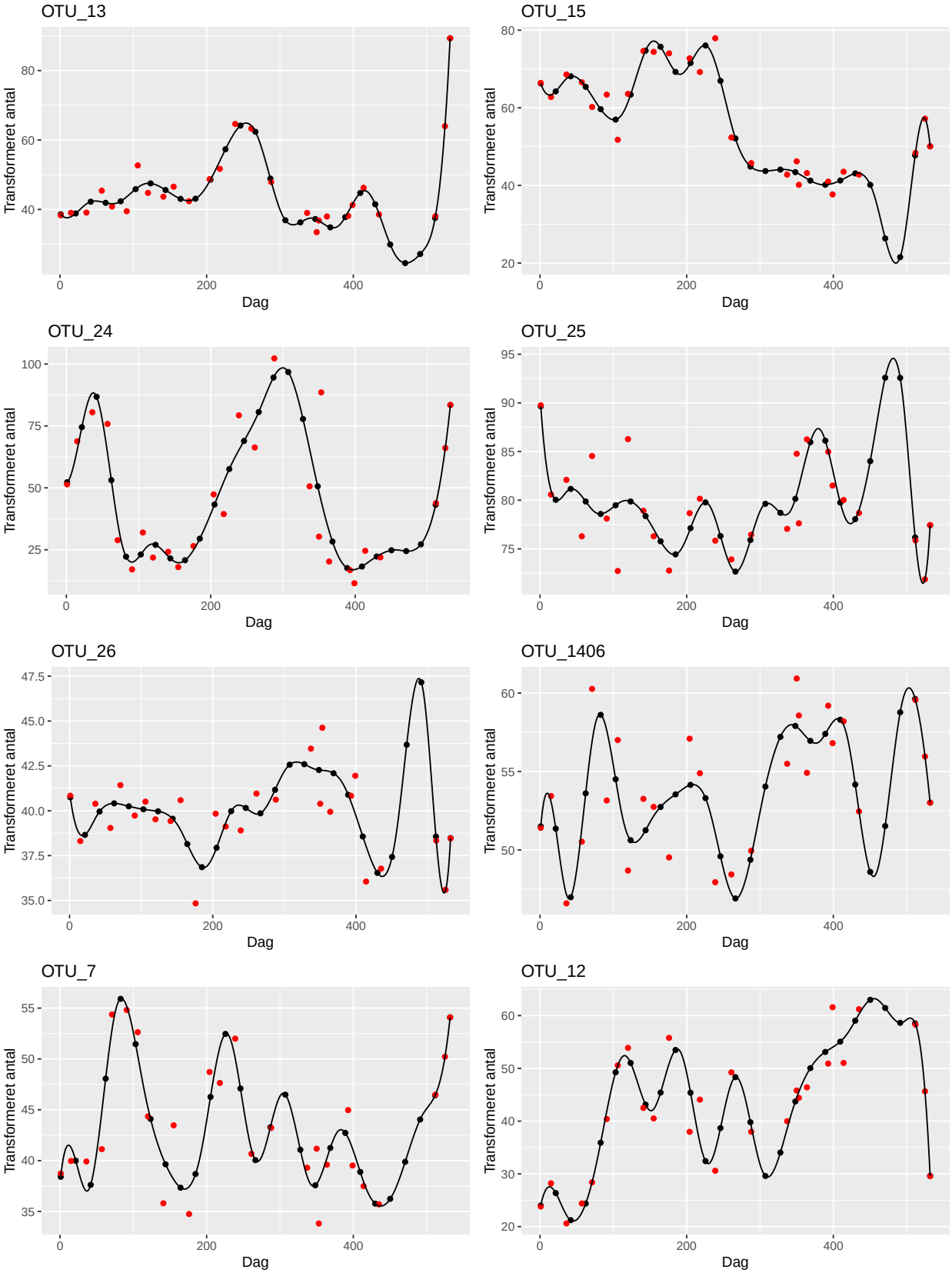


Figur 20: Heatmap over prøverne taget på prøvetagningsdag 8-14 i reaktor T17.

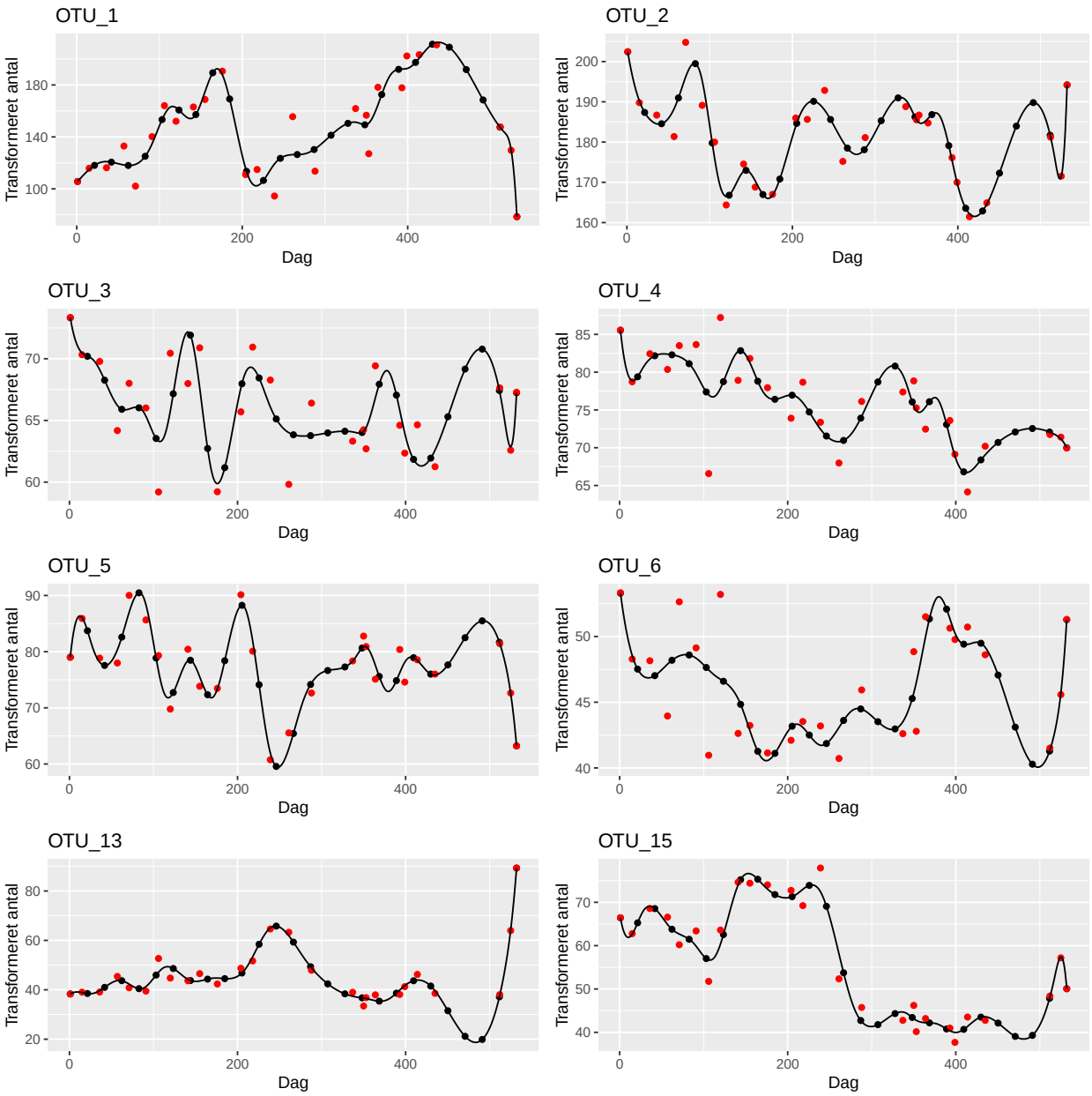
C Splines



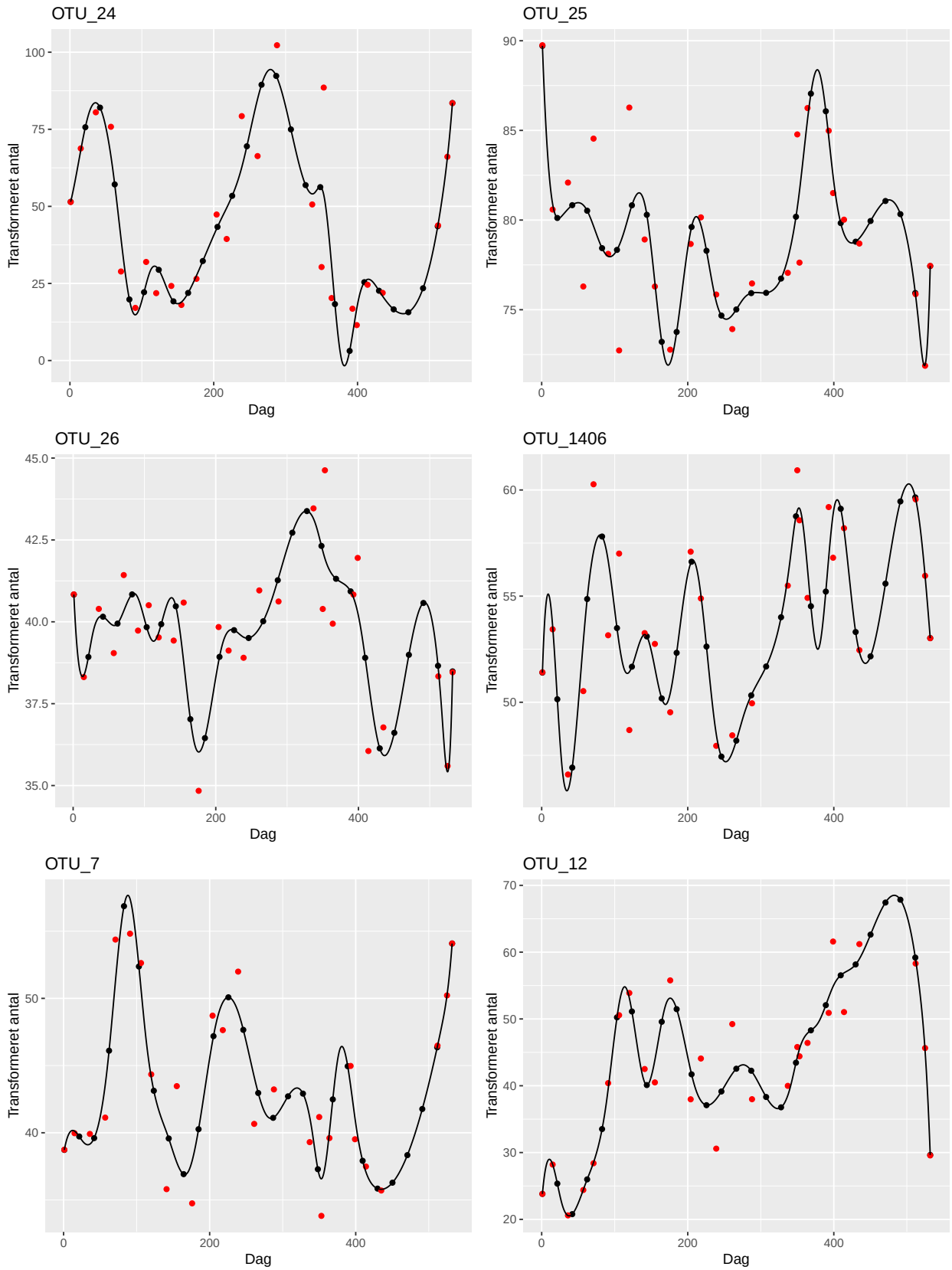
Figur 21: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T14. Knuderne er valgt til at være 15 knuder fordelt ligeligt mellem mindste og største observation



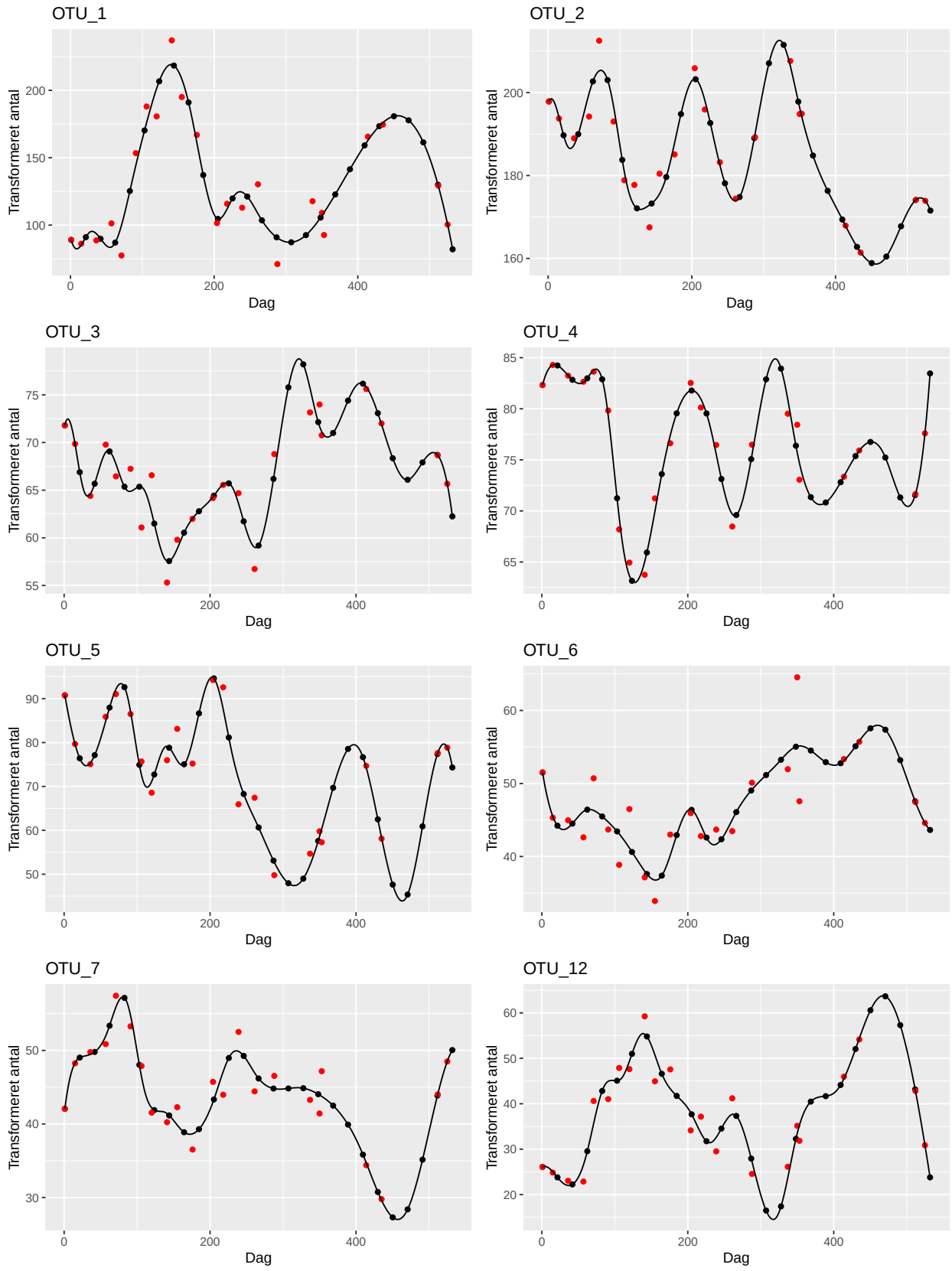
Figur 22: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T14. Knuderne er valgt til at være 15 knuder fordelt ligeligt mellem mindste og største observation



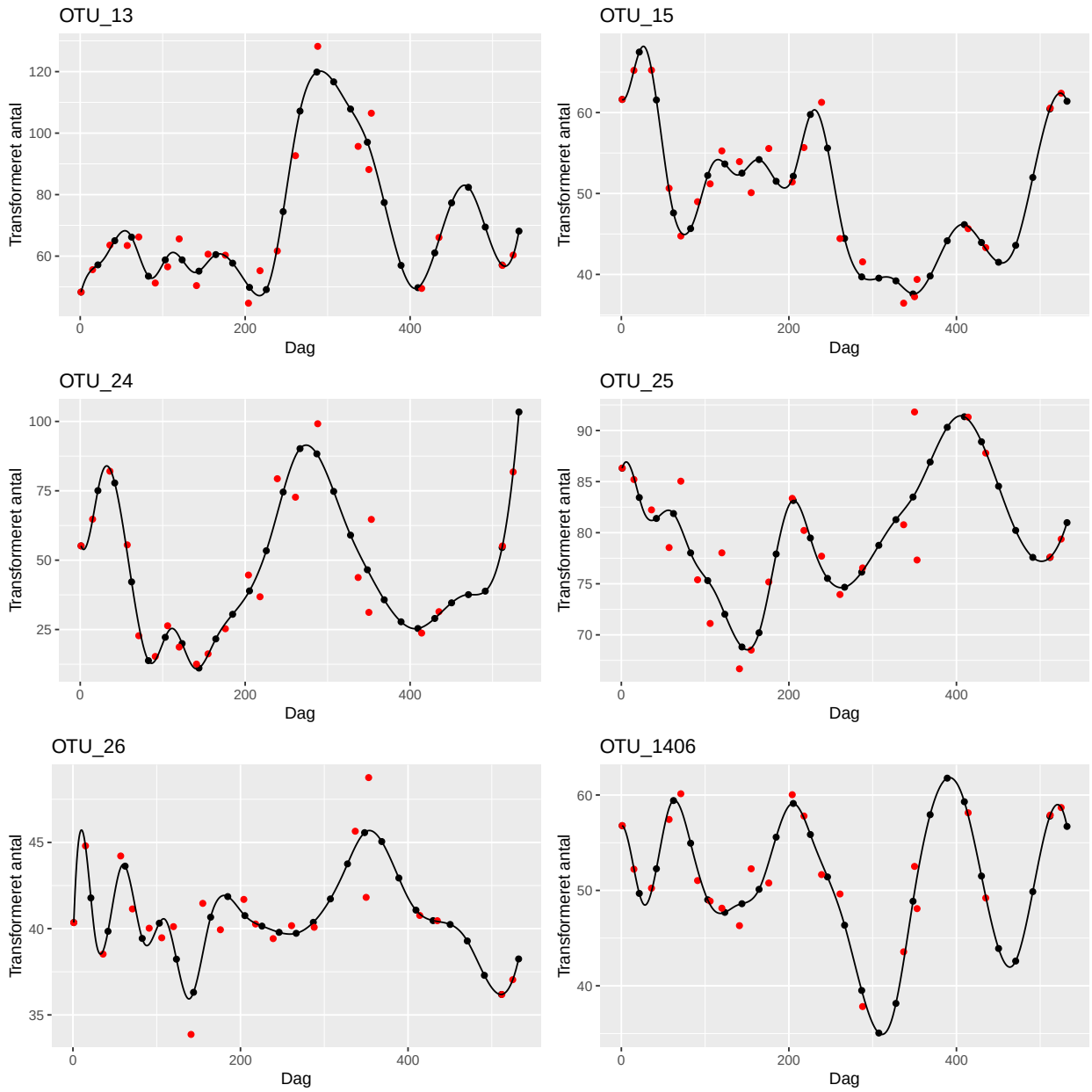
Figur 23: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T14. Knuderne er valgt til percentiler for $p \in \left\{\frac{1}{16}, \frac{2}{16}, \ldots, \frac{15}{16}\right\}$.



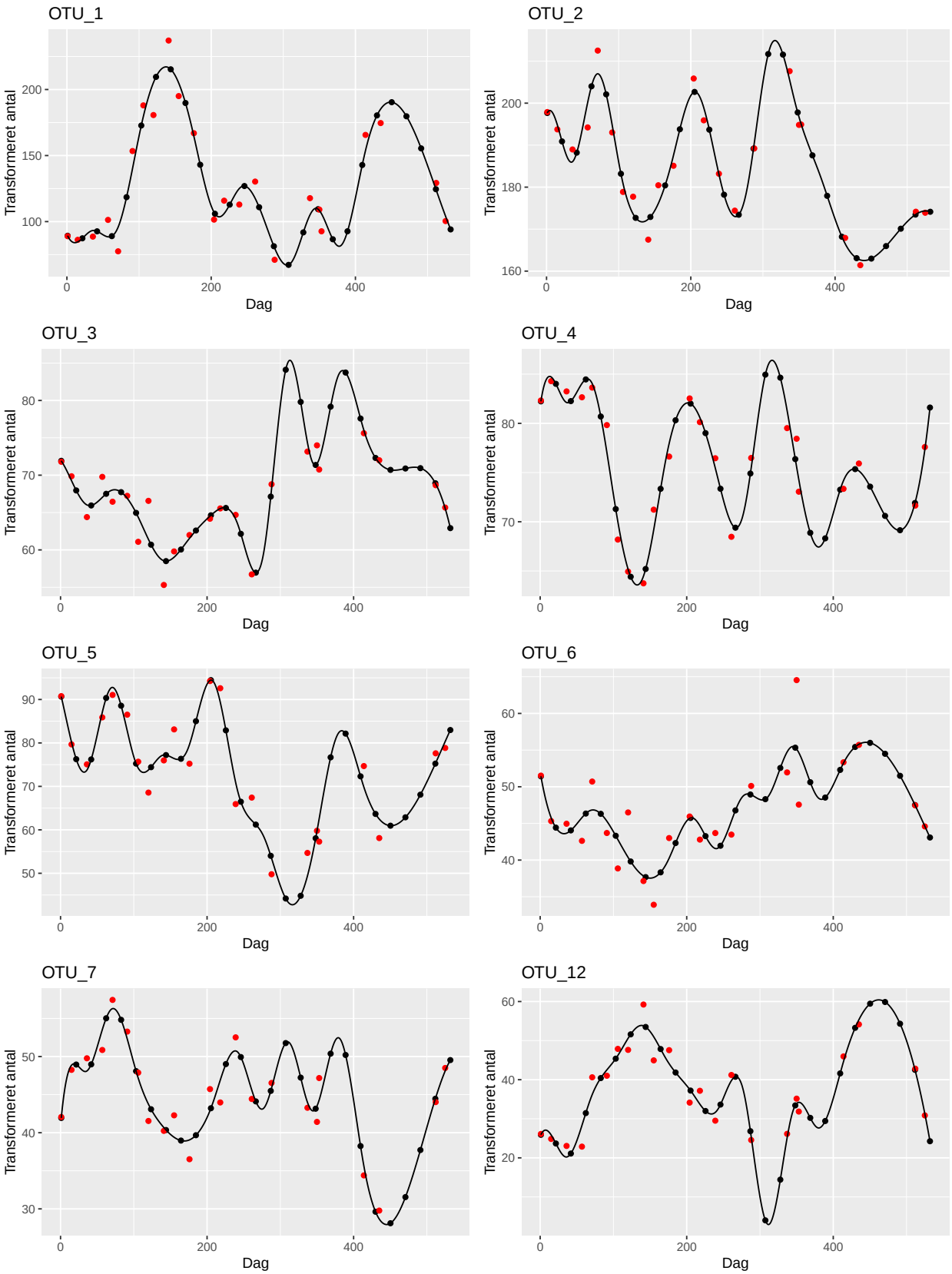
Figur 24: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T14. Knuderne er valgt til percentiler for $p \in \left\{\frac{1}{16}, \frac{2}{16}, \ldots, \frac{15}{16}\right\}$.



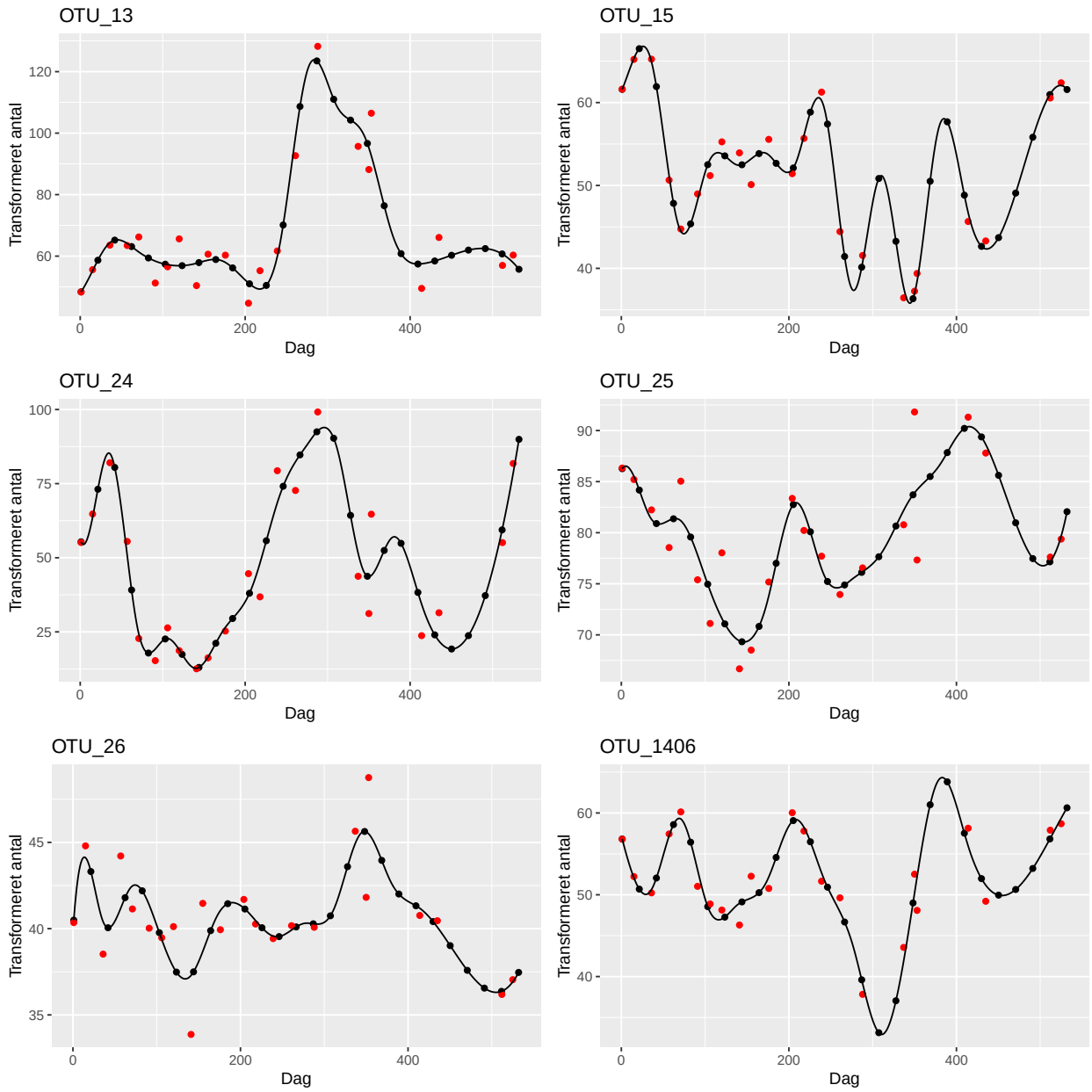
Figur 25: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T16. Knuderne er valgt til percentiler for $p \in \left\{\frac{1}{16}, \frac{2}{14}, \ldots, \frac{13}{14}\right\}$.



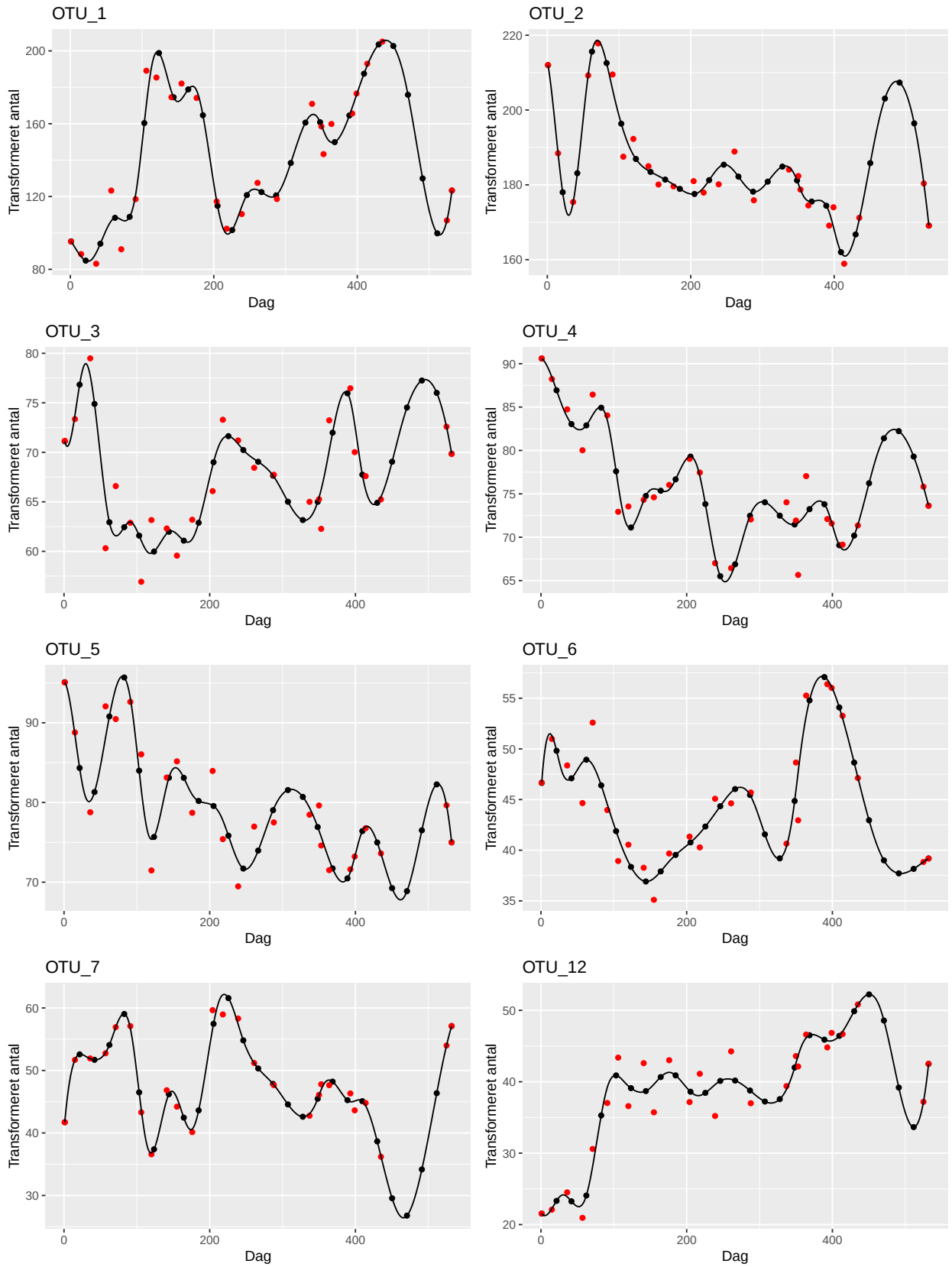
Figur 26: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T16. Knuderne er valgt til percentiler for $p \in \left\{\frac{1}{16}, \frac{2}{14}, \ldots, \frac{13}{14}\right\}$.



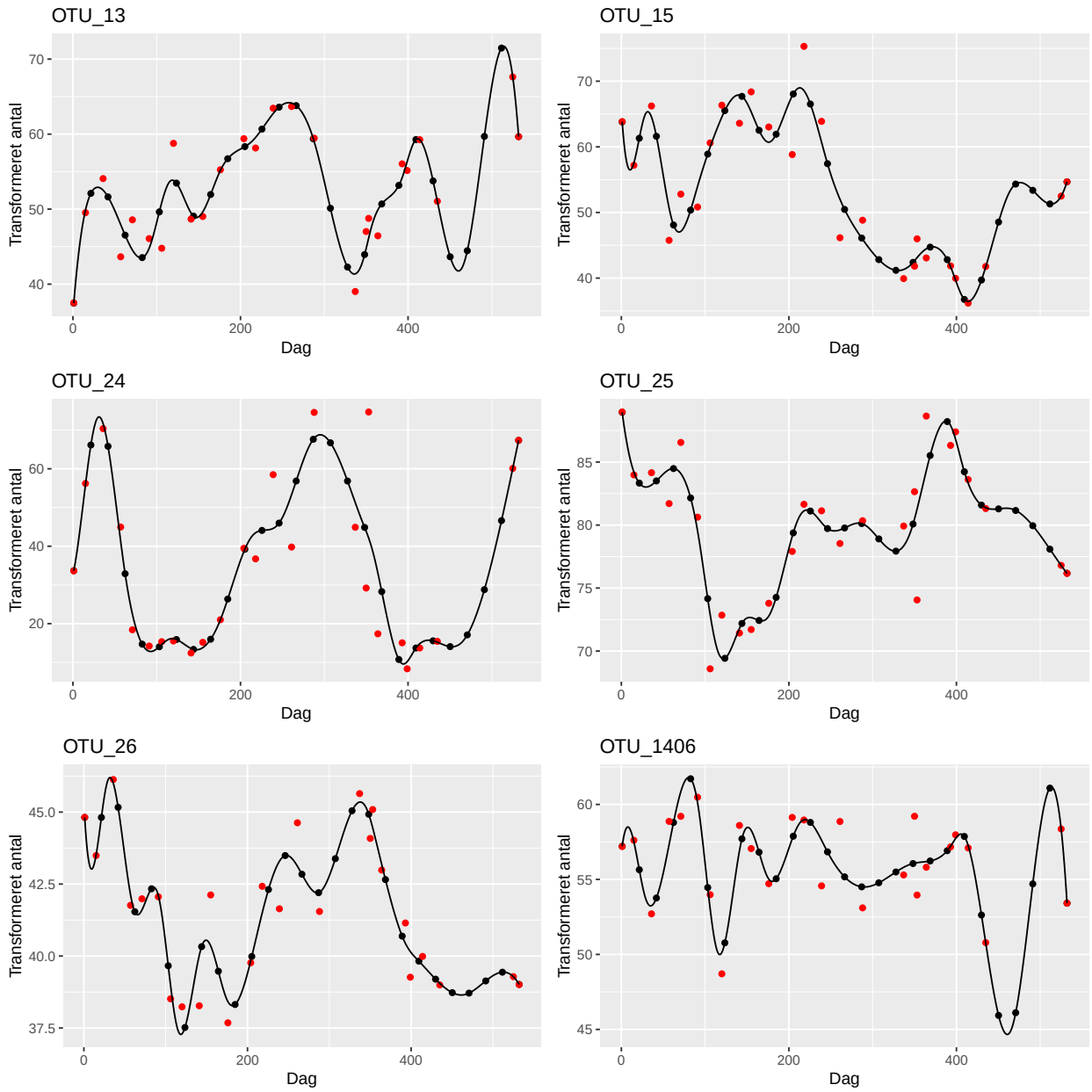
Figur 27: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T16. Knuderne er valgt til at være 13 knuder fordelt ligeligt mellem mindste og største observation.



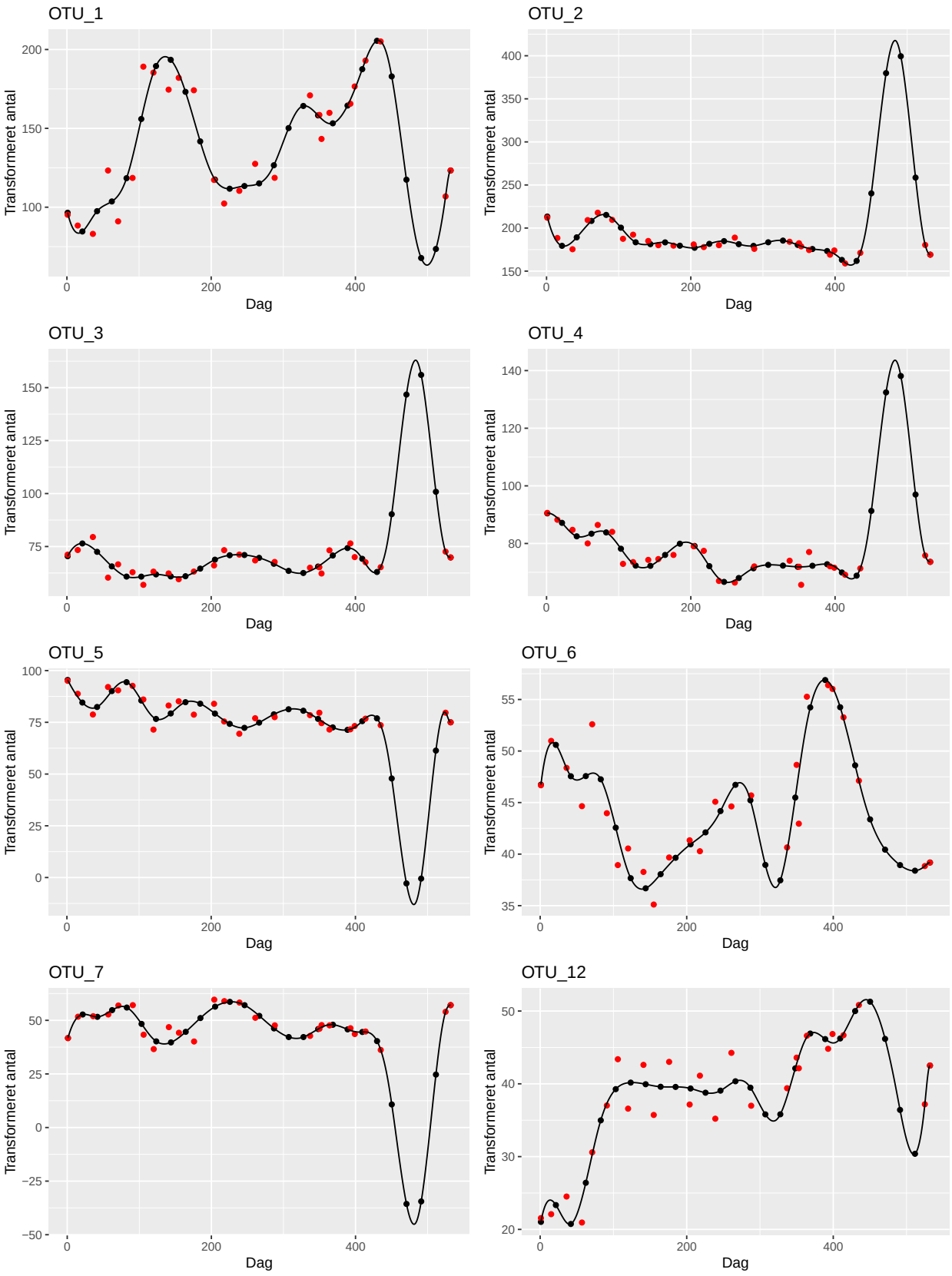
Figur 28: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T16. Knuderne er valgt til at være 13 knuder fordelt ligeligt mellem mindste og største observation.



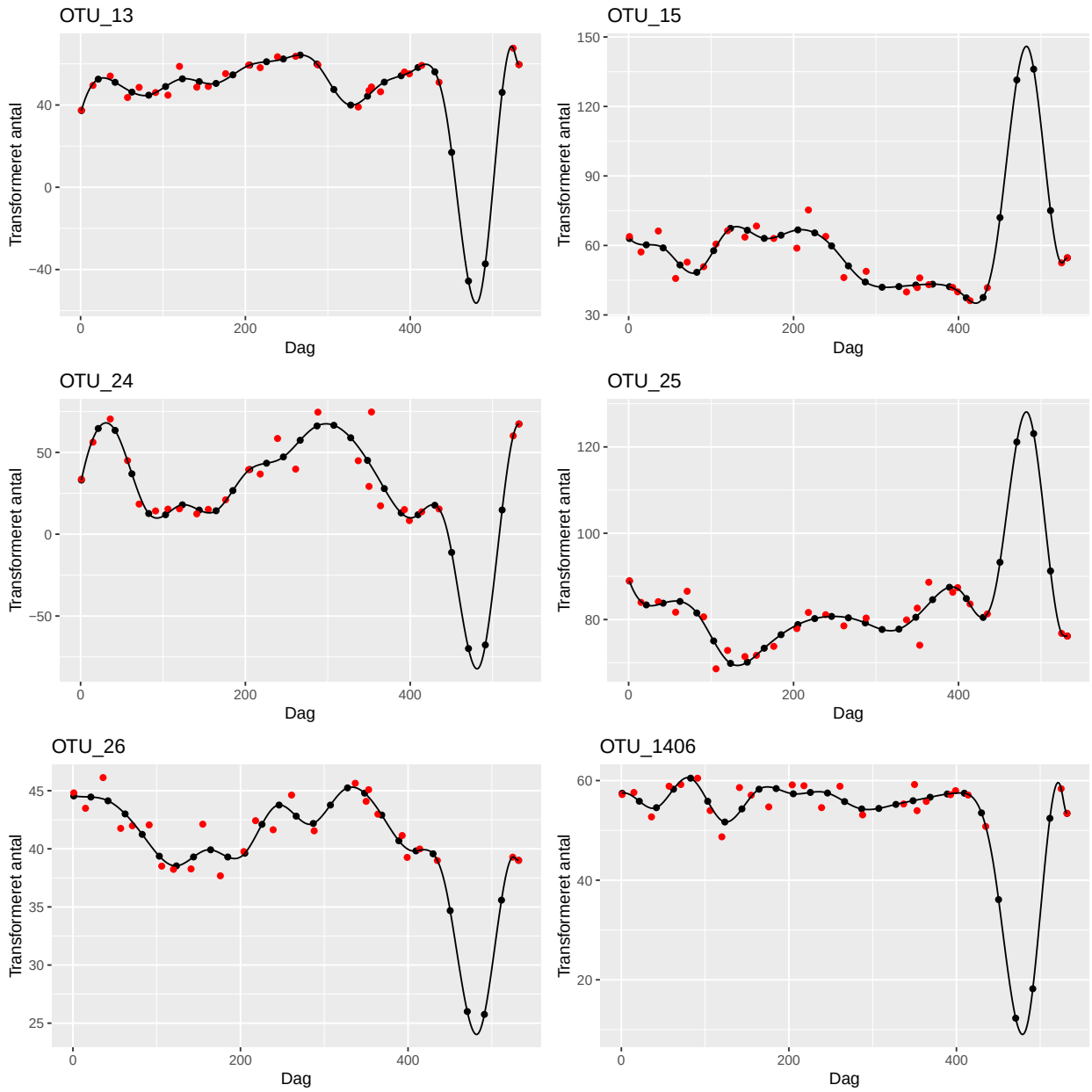
Figur 29: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T17. Knuderne er valgt til percentiler for $p \in \left\{\frac{1}{15}, \frac{2}{15}, \ldots, \frac{14}{15}\right\}$.



Figur 30: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T17. Knuderne er valgt til percentiler for $p \in \left\{\frac{1}{15}, \frac{2}{15}, \ldots, \frac{14}{15}\right\}$.



Figur 31: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T17. Knuderne er valgt til at være 14 knuder fordelt ligeligt mellem mindste og største observation.



Figur 32: Kubisk spline for en delmængde af OTU'er henover tiden taget i reaktor T17. Knuderne er valgt til at være 14 knuder fordelt ligeligt mellem mindste og største observation.

D Funktioner

I dette kapitel vises funktionerne som er implementeret i R med en tilhørende beskrivelse. Funktionen `ptest(x)`, beregner p-værdier for elementerne i `x`, baseret på en to-halet normalfordeling. Funktionen er derfor anvendelig, når p-værdier for flere to-halet tests skal udregnes.

```
ptest=function(x){
  if(!is.vector(x)|!is.numeric(x)){
    stop("Input should be a numeric vector")
    }
2*pnorm(-abs(x))
}
```

Funktionen `modellering(data,B,x)` fitter en AR(1) med et antal forklarende variable, som er signifikante. Er der mindst en regressionskoefficient (med undtagelse af skæring) som ikke er signifikant (signifikansniveau på 5% er anvendt) forskellig fra nul, sorteres den forklarende variabel med størst p-værdi fra. Bemærk, at det kun er de forklarende variable som kan sorteres fra, så selvom den fittede model viser, at lag en ikke er signifikant, sorterer funktionen istedet den forklarende variabel med størst p-værdi fra. Følger modellen ikke en AR(1) vil det ses i outputtet, da den sidste fittede model vil være AR(1) med en ikke signifikant regressionskoefficient for lag en. Funktionen giver et output i form af en liste, med alle de fittede modeller og de tilhørende p-værdier for regressionskoefficienterne i en liste. For at anvende funktionen skal der angives tre oplysninger: `data`, `B` og `ptest`. `data` er et datasæt, gemt som en $n \times m$ matrice, bestående af variablene hen ad søjlerne og observationerne til ækvidistante tidspunkter hen ad rækkerne. `B` består af en $m \times m$ matrix som angiver interaktionerne mellem variablene, hvor 0 står for ingen interaktion og indgange som er forskellig fra nul vil indgå som en forklarende variabel. `x` angiver søjlenummeret, i `data`, for den variabel, der skal modelleres (bedre formulering).
Bemærk, at `ptest` indgår i implementeringen af `modellering(data,B,x)`, hvorfor `ptest` først skal defineres for at `modellering(data,B,x)` virker.

```
modellering=function(data,B,x){
  if(!is.double(x)){stop("not a value")}
  if(!is.matrix(data)){stop("not matrix")}
  if(!is.matrix(B)){stop("not matrix")}
  diag(B)=0
  omgang=(1:14)[B[x,]!=0]
  model=arima(data[,x], c(1,0,0), xreg=data[c(NA,1:26),omgang])
i=1
vec=list()
vec[[i+1]]=test1=ptest(model$coef/sqrt(diag(model$var.coef)))
test1=as.vector(test1)
vec[[i]]=model
while((length(test1)>2 & sum(test1[c(1,3:length(test1))]<0.05)<length(test1)-1)=="TRUE"){
  i=i+2
  maxp=which(test1[1:length(test1)]==max(test1[3:length(test1)]))
  omgang=omgang[-(maxp-2)]
  model=arima(data[,x], c(1,0,0), xreg=data[c(NA,1:26),omgang])
vec[[i+1]]=test1=ptest(model$coef/sqrt(diag(model$var.coef)))
test1=as.vector(test1)
vec[[i]]=model
}
vec
}
```

E Modellering af sparse VAR(1)

I nærværende kapitel fittes sparse VAR(1) for alle tre reaktorer. Da det kun er VAR(1) for reaktor $T14$, som anvendes til dataanalysen er det kun for denne model, der undersøges om residualerne opfylder normalitet, stationaritet og om residualprocessen er seriel ukorreleret. Ved at anvende funktionen `fitVAR()` udføres krydsvalidering for at finde et estimat for lasso-straffen, når denne er fundet estimeres parametrene for VAR(1). I følgende out-put ses det at lasso-straffen varierer meget, og dermed varierer estimaterne for VAR(1) også.

```
load("OTU_spline.Rda")
load("Akvivalent_tider.Rda")
data_samlet=as.data.frame(t(OTU_spline))
data_samlet[] <- lapply(data_samlet, as.character)
colnames(data_samlet) <- data_samlet[1, ]
data_samlet <- data_samlet[-1 ,]
T14_data=data_samlet[1:27,]
T14_data_matrix=data.matrix(T14_data, rownames.force = NA)
T16_data=data_samlet[28:54,]
T16_data_matrix=data.matrix(T16_data, rownames.force = NA)
T17_data=data_samlet[55:81,]
T17_data_matrix=data.matrix(T17_data, rownames.force = NA)
lampda=vector()
diag=vector()
for(x in 1:40){set.seed(x)
  T14_fit=fitVAR(T14_data_matrix, p=1)
tmp=T14_fit$lambda
tmp1=sum(diag(T14_fit$A[[1]])==0)
lampda=c(lampda,tmp[1])
diag=c(diag,tmp1[1])}
lampda; diag
```

```
[1] 0.7380110 0.5086824 0.3847996 0.6724482 0.4634924 0.4223170 0.5086824 0.5582783
 [9] 0.3847996 0.4223170 0.2201965 0.6127097 0.4634924 0.5086824 0.6127097 0.4634924
[17] 0.3847996 0.5086824 0.4223170 0.4223170 0.5086824 0.4223170 0.5582783 0.4634924
[25] 0.2652274 0.5582783 0.5582783 0.3847996 0.4223170 0.3847996 0.6127097 0.4223170
[33] 0.3847996 0.3847996 0.6127097 0.4634924 0.5086824 0.6127097 0.5582783 0.0412608
[41] 0.0412608
 [1] 6 5 4 6 5 5 5 5 4 5 2 6 5 5 6 5 4 5 5 5 5 5 5 5 2 5 5 4 5 4 6 5 4 4 6 5 5 6 5 3
```

```
library(sparsevar)
set.seed(11) # 2 nuller i diagonalen,  lampda=0.2201965
T14_fit=fitVAR(T14_data_matrix, p=1)
T14_B=T14_fit$A[[1]]
T14_fit$lambda; diag(T14_fit$A[[1]]); T14_fit$A
```

```
[1] 0.2201965
 [1] 0.17764129 0.07183886 0.00000000 0.06781112 0.36354260 0.34330169 0.89960228
 [8] 0.55734649 0.53908805 0.49428956 0.34502379 0.00000000 0.08314770 0.81059709
[[1]]
            [,1]        [,2]       [,3]        [,4]       [,5]        [,6]         [,7]
 [1,]  0.1776413  0.00000000  0.0000000  0.00000000  0.0000000  0.00000000  0.000000000
 [2,]  0.1857081  0.07183886  0.0000000  0.00000000  0.0000000  0.00000000  0.139238454
 [3,]  0.0000000  0.00000000  0.0000000 -0.31033306  0.1400266  0.00000000  0.000000000
 [4,]  0.0000000  0.00000000  0.0000000  0.06781112  0.0000000  0.00000000  0.000000000
 [5,]  0.0000000  0.00000000  0.0000000  0.00000000  0.3635426  0.00000000  0.000000000
 [6,] -0.6127198  0.00000000  0.2230034  0.00000000 -0.3957002  0.34330169  0.000000000
 [7,]  6.1629206 -0.07633670  0.0000000  0.00000000  2.6034070 -0.01117851  0.899602280
 [8,] -0.7174985  0.00000000  0.0000000  0.39769726 -0.8039343  0.00000000 -0.055886175
 [9,]  0.0000000  0.00000000  0.0000000  0.00000000  0.0000000  0.00000000  0.000000000
[10,] -1.6577241  0.00000000 -0.6495595  2.33256384 -0.5167425  0.29516546 -0.209938017
[11,]  0.0000000  0.00000000  0.1810734  0.00000000 -0.1449446  0.00000000 -0.028699376
[12,]  0.0000000  0.00000000  0.0000000  0.00000000  0.0000000  0.00000000  0.001039081
[13,]  0.0000000  0.00000000  0.0000000  0.00000000  0.1468354  0.00000000  0.000000000
[14,]  0.0000000  0.00000000  0.2641461  0.15056230  0.0000000 -0.12192921  0.000000000
```

```
              [,8]        [,9]         [,10]       [,11]        [,12]       [,13]
 [1,]  0.000000000  0.0000000  0.005394777  0.0000000  0.00000000  0.0000000
 [2,]  0.000000000 -0.4622357 -0.155189862  0.0000000  0.00000000  0.0000000
 [3,] -0.246296743  0.3518191  0.000000000 -0.2196021  0.00000000  0.0000000
 [4,] -0.010041509  0.0000000  0.000000000  0.0000000  0.00000000  0.0000000
 [5,]  0.000000000  0.0000000  0.000000000  0.0000000  0.05562825  0.0000000
 [6,] -0.145974957  0.0000000  0.166161335  0.0000000  0.00000000  0.0000000
 [7,]  0.754723597  0.0000000 -0.167051664 -0.4752555  0.00000000  0.0000000
 [8,]  0.557346492  0.0000000  0.000000000  0.5891046  0.97059421  0.0000000
 [9,]  0.000000000  0.5390880  0.022252112  0.0000000 -0.04610656  0.0000000
[10,]  0.000000000 -1.1060015  0.494289560  0.0000000  0.00000000 -1.2632757
[11,]  0.000000000  0.0000000  0.000000000  0.3450238  0.00000000  0.0000000
[12,] -0.132846790  0.0000000  0.000000000  0.0000000  0.00000000  0.0000000
[13,] -0.005882827  0.0000000  0.000000000  0.0000000  0.00000000  0.0831477
[14,]  0.000000000  0.2144117  0.000000000  0.0000000  0.00000000  0.0000000
             [,14]
 [1,]  0.00000000
 [2,] -0.05952616
 [3,]  0.00000000
 [4,]  0.00000000
 [5,]  0.00000000
 [6,]  0.00000000
 [7,]  0.00000000
 [8,]  0.02560487
 [9,]  0.00000000
[10,]  0.00000000
[11,]  0.00000000
[12,] -0.02766429
[13,] -0.02161781
[14,]  0.81059709
```

Der fittes en sparse VAR(1) for OTU'erne i reaktor $T16$.

```
set.seed(2) # ingen nuller i diagonalen, lampda=0.1469233
T16_fit=fitVAR(T16_data_matrix, p=1)
T16_fit$lambda; diag(T16_fit$A[[1]]);T16$A
```

```
[1] 0.1469233
 [1] 0.24387466 0.30591116 0.03827453 0.31120957 0.08297630 0.42967071 0.82816201
 [8] 0.00000000 0.38450978 0.83384328 0.31291648 0.48047950 0.51140797 0.68734391
[[1]]
             [,1]         [,2]         [,3]          [,4]         [,5]         [,6]
 [1,]  0.2438747 -0.05965998  0.00000000  0.0000000000  0.00000000  0.00000000
 [2,]  0.9519614  0.30591116 -0.01922145  0.0000000000  0.00000000 -0.12618586
 [3,]  1.7202024  0.00000000  0.03827453  0.0000000000  0.00000000  0.14364661
 [4,]  0.0000000  0.00000000  0.00000000  0.3112095688  0.00000000  0.00000000
 [5,]  0.0000000  0.00000000 -0.15965078  0.0000000000  0.08297630  0.00000000
 [6,]  0.0000000 -0.03012469  0.01784248 -0.0698629663 -0.99505301  0.42967071
 [7,]  0.0000000  0.22801418  0.00000000 -0.8828041675  0.00000000  0.00000000
 [8,] -1.3189838  0.00000000 -0.21236524  0.0000000000 -0.62631160 -0.06346549
 [9,] -0.0974900 -0.05407883  0.06177082  0.0000000000  0.00000000  0.00000000
[10,] -3.6113687  0.00000000  0.00000000  0.0001168627  0.07404807  0.00000000
[11,]  0.0000000  0.00000000  0.01133480  0.0000000000  0.00000000  0.04624288
[12,]  0.5561907  0.00000000  0.00000000  0.0000000000  0.00000000  0.00000000
[13,]  0.4898096  0.00000000  0.00000000  0.1402393609  0.00000000  0.00000000
[14,] -0.4412033  0.05502961  0.07793765  0.0000000000  0.20357158  0.03144107
             [,7]          [,8]         [,9]        [,10]        [,11]       [,12]
 [1,]  0.000000000 0.0001072010  0.00000000  0.00000000  0.00000000  0.0000000
 [2,]  0.087323741 0.0000000000  0.08994132 -0.16644645  0.00000000  0.0000000
 [3,]  0.000000000 0.0000000000  0.00000000 -0.13818891  0.00000000  0.9024650
 [4,] -0.005596696 0.0725240493  0.00000000  0.00000000  0.00000000  0.0000000
 [5,]  0.000000000 0.0008635286  0.00000000  0.00000000  0.00000000  0.0000000
 [6,] -0.082170308 0.2293146763  0.43708045  0.00000000 -0.36846895  0.0000000
 [7,]  0.828162014 0.0000000000 -0.72709094 -0.69793219  3.09118641 -2.0857920
 [8,]  0.000000000 0.0000000000  0.00000000  0.61661550  0.05836631 -0.4435042
 [9,]  0.000000000 0.0000000000  0.38450978  0.07017225 -0.14633477  0.0000000
[10,] -0.092878699 0.0000000000  0.00000000  0.83384328 -0.52674057  0.5326637
[11,] -0.071569130 0.0000000000  0.00000000  0.00000000  0.31291648  0.0000000
[12,]  0.000000000 0.0000000000  0.00000000 -0.04365196  0.00000000  0.4804795
[13,]  0.000000000 0.0000000000  0.00000000  0.00000000 -0.01454168  0.0000000
[14,]  0.000000000 0.0000000000  0.00000000 -0.04770181  0.00000000  0.0000000
             [,13]         [,14]
 [1,]  0.00000000 -0.008057883
 [2,]  0.21573644 -0.076635889
 [3,] -0.83198764  0.618880701
 [4,]  0.21775022 -0.093037859
 [5,]  0.46732404 -0.160773084
 [6,] -0.20028222  0.000000000
 [7,]  4.18055975 -0.349666640
 [8,]  0.00000000 -1.138922146
 [9,]  0.00000000  0.000000000
[10,] -0.04384692  0.304457804
[11,] -0.22947444  0.227725977
[12,]  0.00000000  0.116907661
```

```
[13,]  0.51140797  0.000000000
[14,]  0.00000000  0.687343909
```

Der fittes en sparse VAR(1) for OTU'erne i reaktor $T17$.

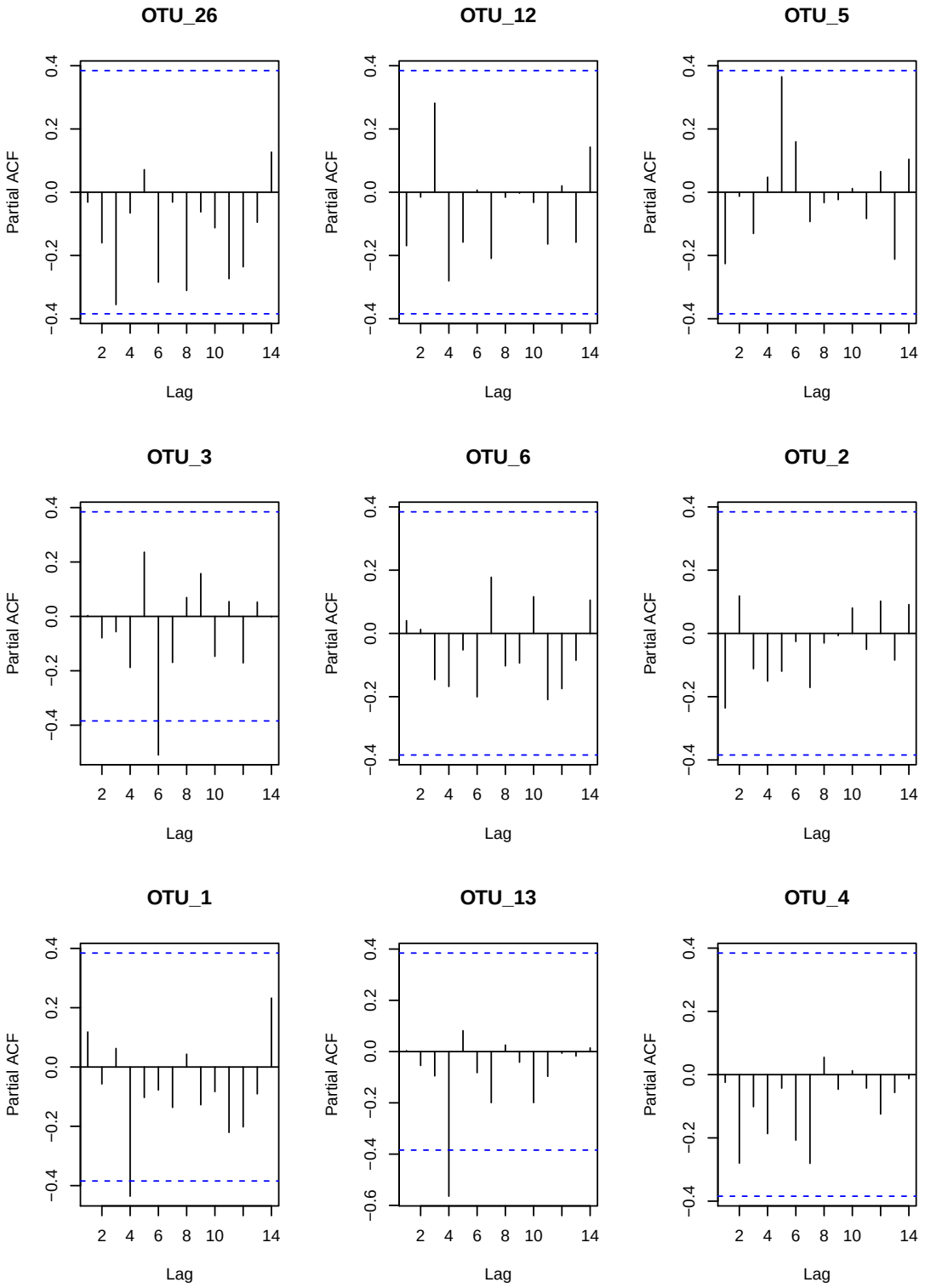
```
set.seed(4) # ingen nuller i diagonalen, lampda=0.1469233
T17_fit=fitVAR(T17_data_matrix, p=1)
T17_fit$lambda; diag(T17_fit$A[[1]]);T17$A
```

```
[1] 0.2136689
 [1] 0.23618891 0.51436456 0.00000000 0.49007139 0.56635771 0.42085881 0.04956529
 [8] 0.34397304 0.54289408 0.52149413 0.05862127 0.04981133 0.53957813 0.62545549
[[1]]
              [,1]          [,2]         [,3]         [,4]          [,5]         [,6]
 [1,]  0.23618891  0.000000000  0.00000000  0.00000000  0.000000000  0.00000000
 [2,]  0.00000000  0.514364564  0.00000000 -0.05038942  0.000000000  0.00000000
 [3,]  0.00000000 -0.454450239  0.00000000 -0.08122392  0.000000000  0.05641360
 [4,]  0.00000000  0.000000000  0.00000000  0.49007139  0.000000000  0.00000000
 [5,]  0.65307712  0.000000000  0.00000000  0.00000000  0.566357712  0.00000000
 [6,]  0.00000000 -0.076076075  0.00000000  0.00000000  0.000000000  0.42085881
 [7,]  0.00000000  0.000000000  0.21691632  0.00000000  0.819724454 -0.69571123
 [8,]  0.00000000  0.126559726  0.00000000  0.41460376 -0.007859249  0.04365038
 [9,]  0.00000000 -0.009130139  0.00000000  0.00000000  0.000000000  0.00000000
[10,]  0.21883641  0.000000000  0.00000000  0.00000000 -0.490011059  0.00000000
[11,]  0.00000000 -0.225990653  0.00000000  0.00000000  0.000000000  0.00000000
[12,]  0.00000000 -0.098473628  0.00000000  0.00000000  0.000000000  0.00000000
[13,]  0.07391453  0.000000000 -0.04132785  0.00000000  0.000000000 -0.02937797
[14,] -0.72054028  0.000000000  0.08252467 -0.21965547  0.000000000  0.00000000
               [,7]         [,8]        [,9]         [,10]        [,11]        [,12]
 [1,]  0.0000000000  0.00000000  0.0000000  0.032044785  0.00000000  0.00000000
 [2,]  0.0000000000  0.00000000 -0.1583086 -0.039919770  0.00000000  0.19700692
 [3,]  0.0000000000  0.00000000  0.0000000  0.000000000  0.00000000  0.00000000
 [4,]  0.0000000000  0.00000000  0.0000000  0.000000000 -0.10165647  0.00000000
 [5,]  0.0000000000  0.00000000  0.0000000  0.000000000  0.00000000  0.00000000
 [6,]  0.0000000000  0.00000000  0.2607095  0.000000000  0.56552862 -2.50292051
 [7,]  0.0495652938 -0.80791131 -0.6564341 -0.844099401  0.24028387  0.00000000
 [8,]  0.0000000000  0.34397304  0.0000000 -0.038135834  0.00000000  0.10609414
 [9,]  0.0000000000 -0.02288748  0.5428941  0.000000000 -0.04313518 -0.22444559
[10,] -0.2321277911  0.58376834  0.0000000  0.521494127 -0.68313421  0.00000000
[11,] -0.0497057571  0.08954228  0.0000000  0.014672482  0.05862127  0.58778969
[12,] -0.0002603475  0.00000000  0.0000000  0.000000000  0.00000000  0.04981133
[13,]  0.0000000000  0.00000000  0.0000000  0.004694791  0.00000000  0.00000000
[14,]  0.0000000000  0.00000000  0.1949854 -0.004788833  0.00000000  0.00000000
            [,13]       [,14]
 [1,]  0.00000000  0.0000000
 [2,]  0.00000000 -0.1418276
 [3,]  0.00000000  0.0000000
 [4,]  0.05226647  0.0000000
 [5,]  0.00000000  0.0000000
 [6,]  0.00000000  0.0000000
 [7,] -2.95887347 -1.6099565
 [8,]  0.00000000  0.1893944
 [9,]  0.00000000  0.0000000
[10,]  0.00000000  0.0171082
[11,]  0.00000000  0.1441719
[12,]  0.00000000  0.0000000
[13,]  0.53957813  0.0000000
[14,] -0.13818891  0.6254555
```

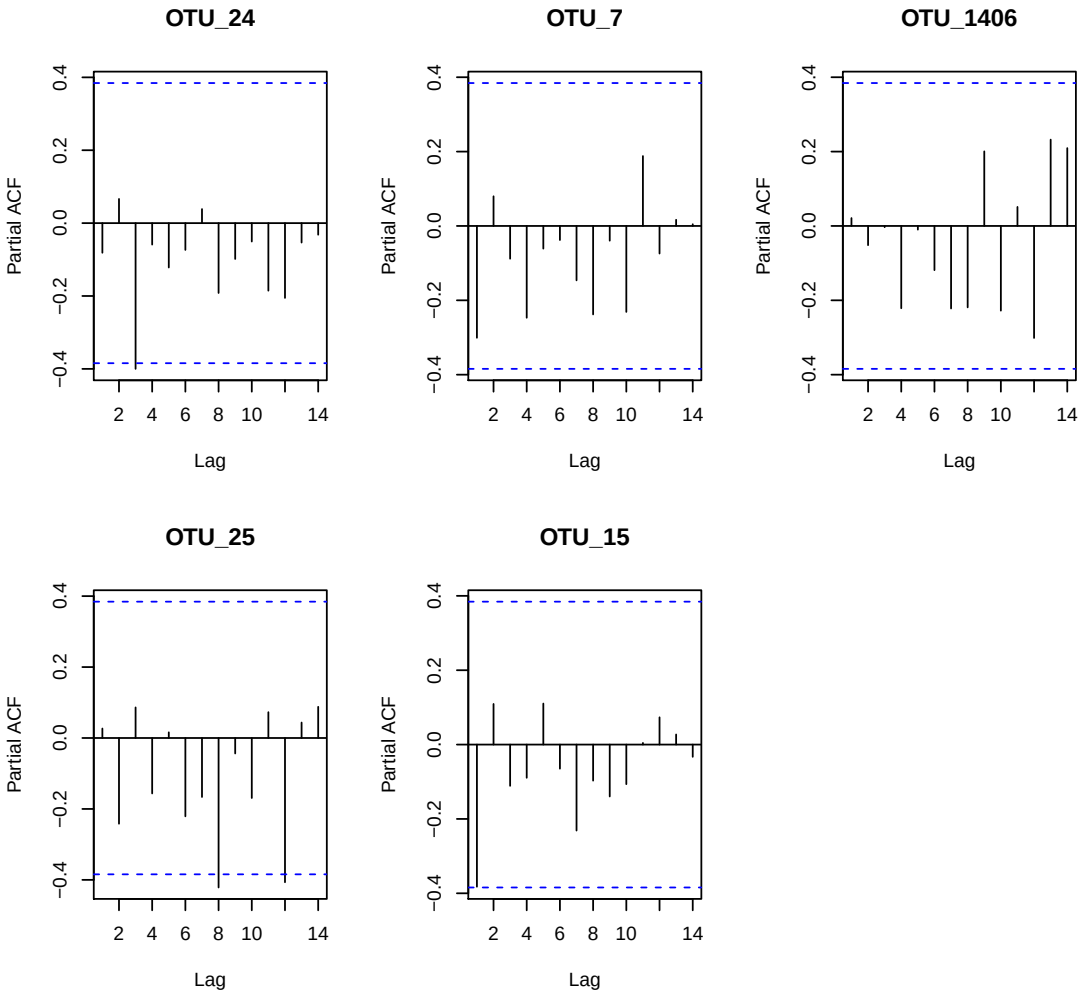
E.1 Modelvalidering af sparse VAR(1) for reaktor T14

Residualerne beregnes for reaktor $T14$ og der undersøges om residualerne er seriel ukorrelleret.

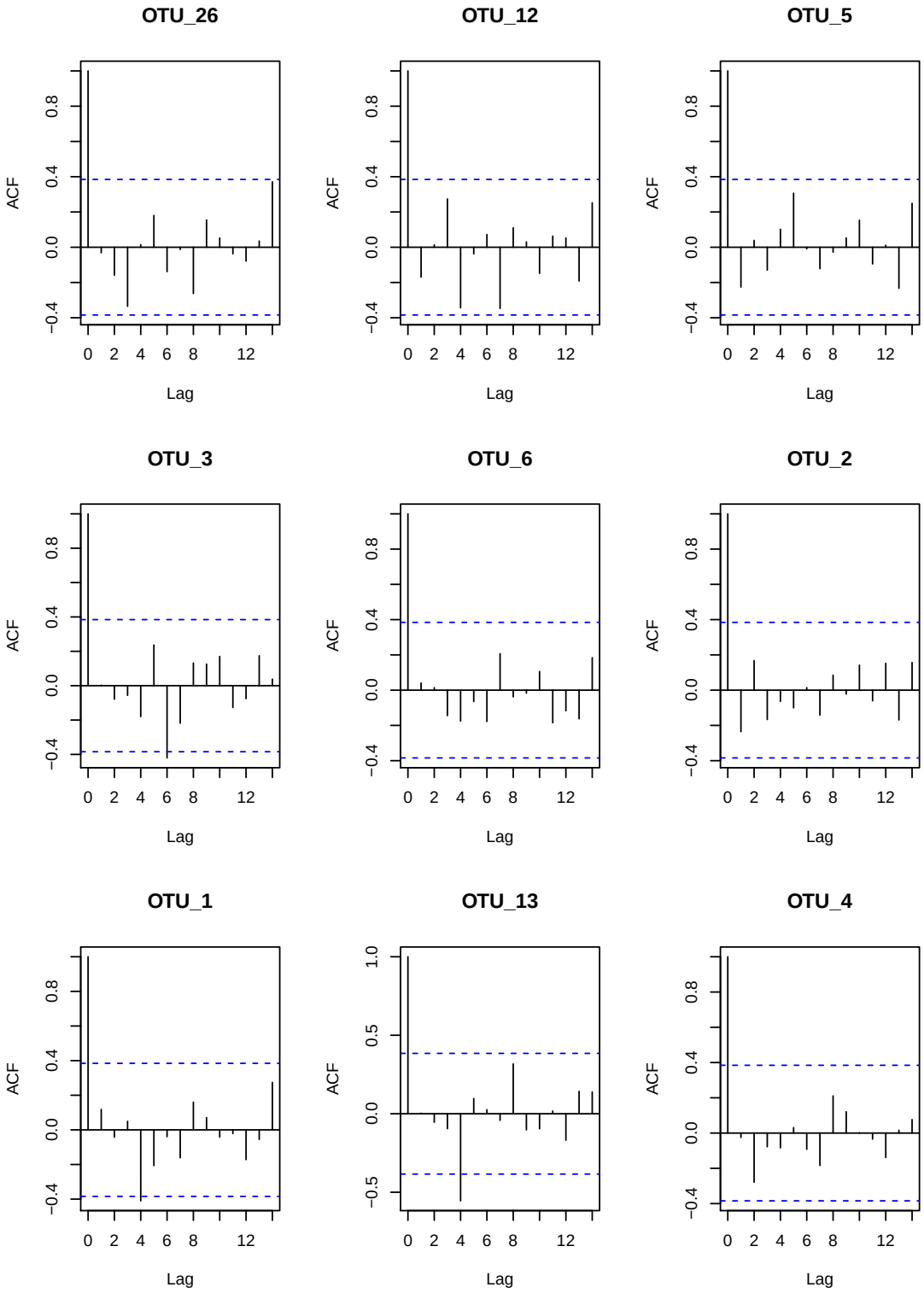
```
output <- matrix(unlist(T14_fit$A), ncol = 14)
output_mean <- matrix(unlist(T14_fit$series), ncol = 14)
res=list()
for(i in 1:26){res[[i]]=-output%*%output_mean[i,]+output_mean[i+1,]}
res=matrix(unlist(res), ncol = 14)
res[,1]
pacflist=list()
for(x in 1:14){tmp=pacf(res[,x])
pacflist[[x]]=tmp}
acflist=list()
for(x in 1:14){tmp=acf(res[,x])
acflist[[x]]=tmp}
```



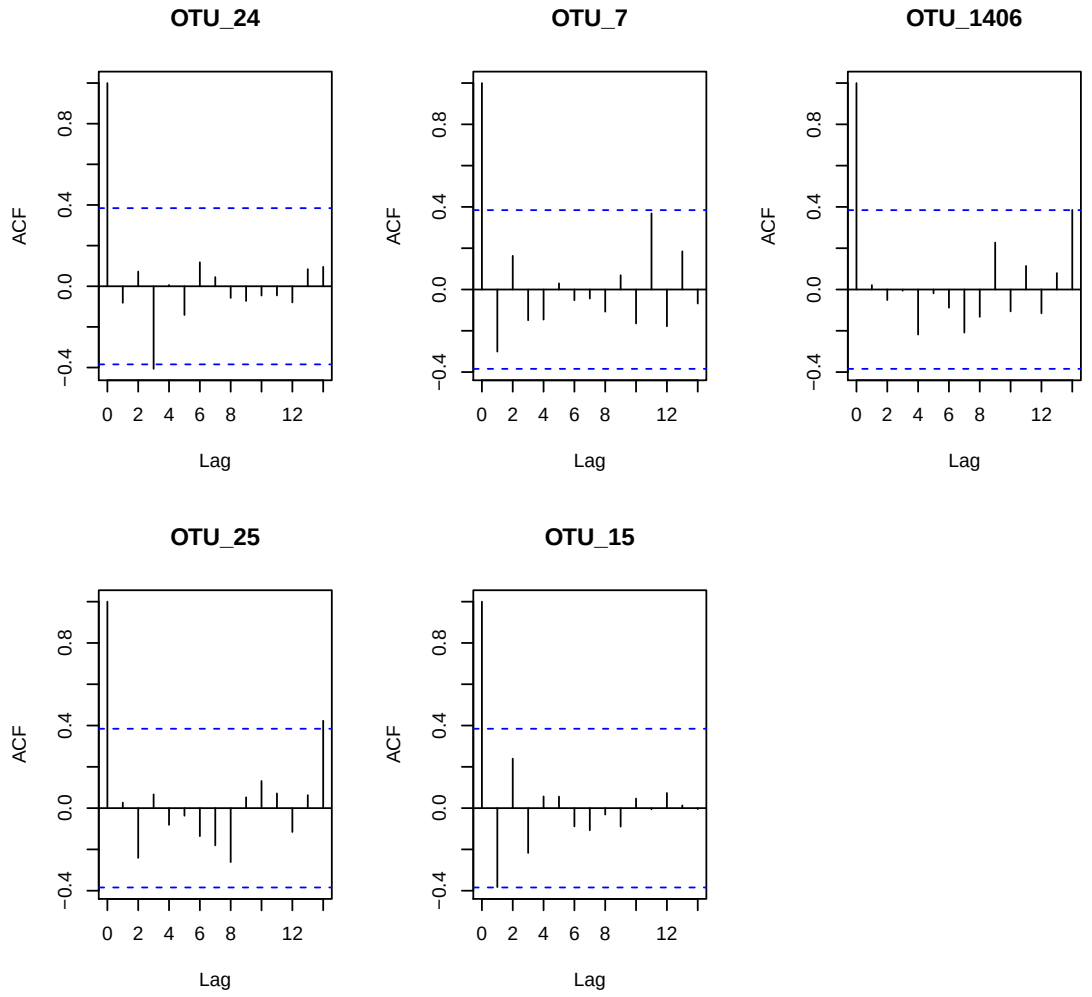
Figur 33: PACF korrelogrammer af residualerne for de første 9 tidsserier i sparse VAR(1).



Figur 34: PACF korrelogrammer af residualerne for de sidste 5 tidsserier i sparse VAR(1).



Figur 35: ACF korrelogrammer af residualerne for de første 9 tidsserier i sparse VAR(1).



Figur 36: ACF korrelogrammer af residualerne for de sidste 5 tidsserier i sparse VAR(1).

Ved hjælp af Box-Pierce testen undersøges om der er en signifikant autokorrelation.

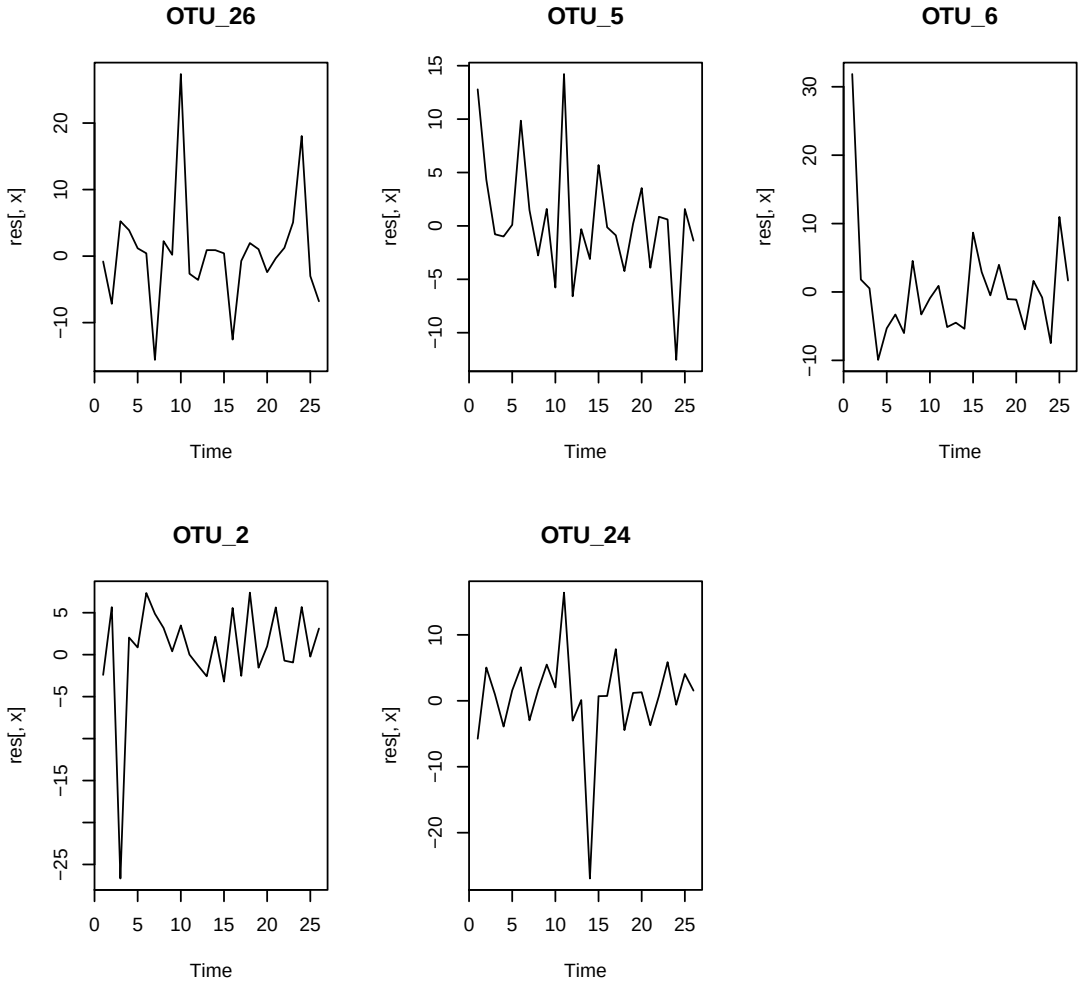
```
library(tseries)
acf_test=list()
for(x in 1:14){tmp=Box.test(res[,x])
  acf_test$alt[[x]]=tmp
  acf_test$p[[x]]=tmp$p.value
}
acf_test$p
```

```
 [1] 0.87237970 0.38847532 0.24953273 0.98791760 0.83687189 0.22920202 0.54686828
 [8] 0.98563411 0.90057866 0.67822828 0.12499036 0.91374545 0.89149065 0.05116456
```

Af Box-Pierce testen kan man ikke forkaste nulhypotesen om, at der er en autokorrelation for mindst en af lagene. Der undersøges om residualerne følger stationaritet:

```
stationaritet_test=list()
for(x in 1:14){tmp=adf.test(res[,x])
  stationaritet_test$alt[[x]]=tmp
  stationaritet_test$p[[x]]=tmp$p.value
}
stationaritet_test$p; names(T14_data)
```

```
[1] 0.01866254 0.99000000 0.01000000 0.33229609 0.02059865 0.01000000 0.33434325
 [8] 0.20687446 0.17172548 0.04446006 0.13267261 0.38646747 0.27650989 0.18087951
 [1] "OTU_26"   "OTU_12"   "OTU_5"    "OTU_3"    "OTU_6"    "OTU_2"    "OTU_1"
 [8] "OTU_13"   "OTU_4"    "OTU_24"   "OTU_7"    "OTU_1406" "OTU_25"   "OTU_15"
```



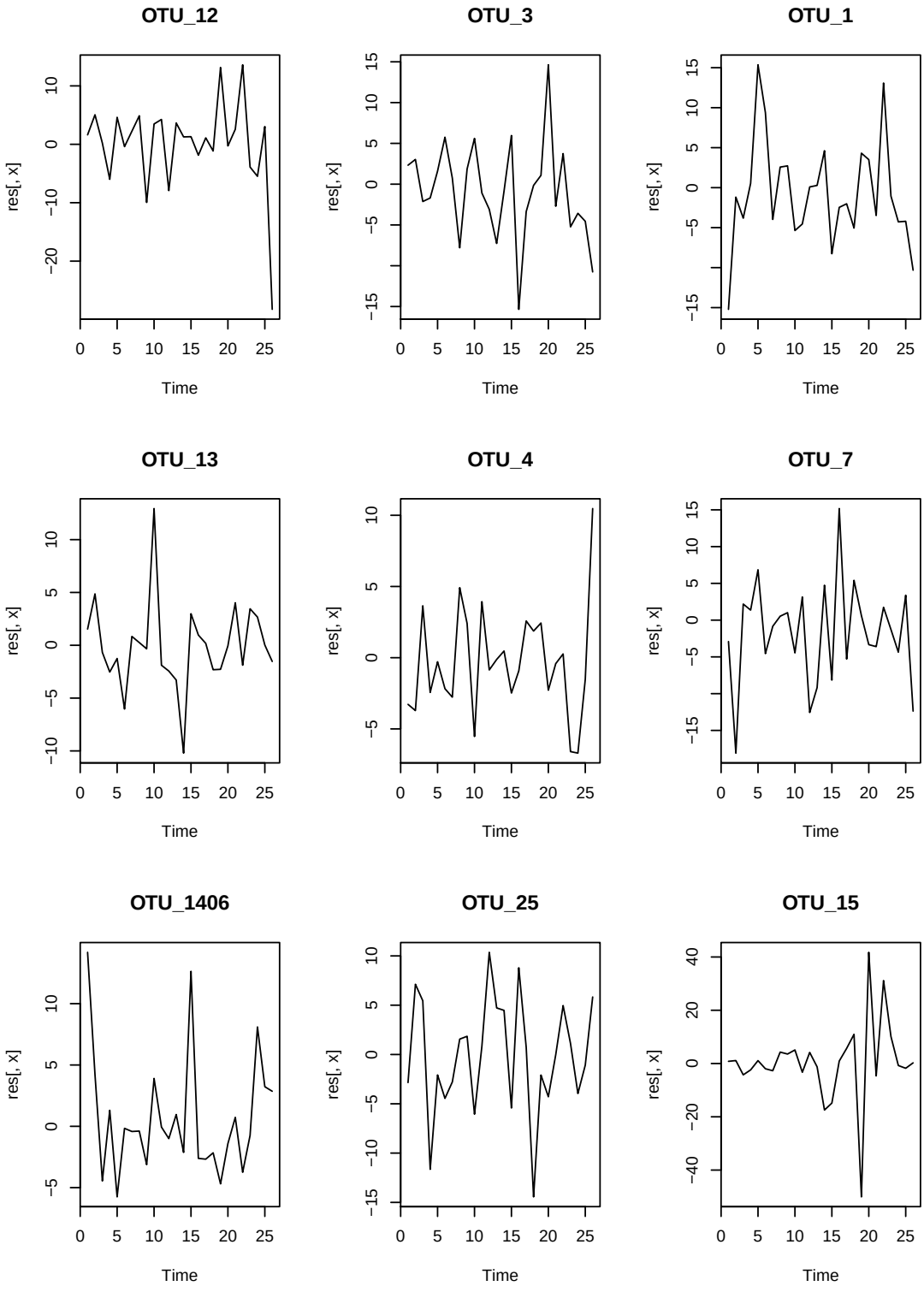
Figur 37: Residualprocessen for de stationære tidsserier i VAR(1).

P-værdien for OTU 12 er ekstrem høj, hvilket virker mærkværdig, da p-værdierne for de øvrige tidsserier ikke viser samme tendens. Der udføres derfor en Dickey-Fuller test for OTU 12 igen, men hvor en af residualerne er fjernet.

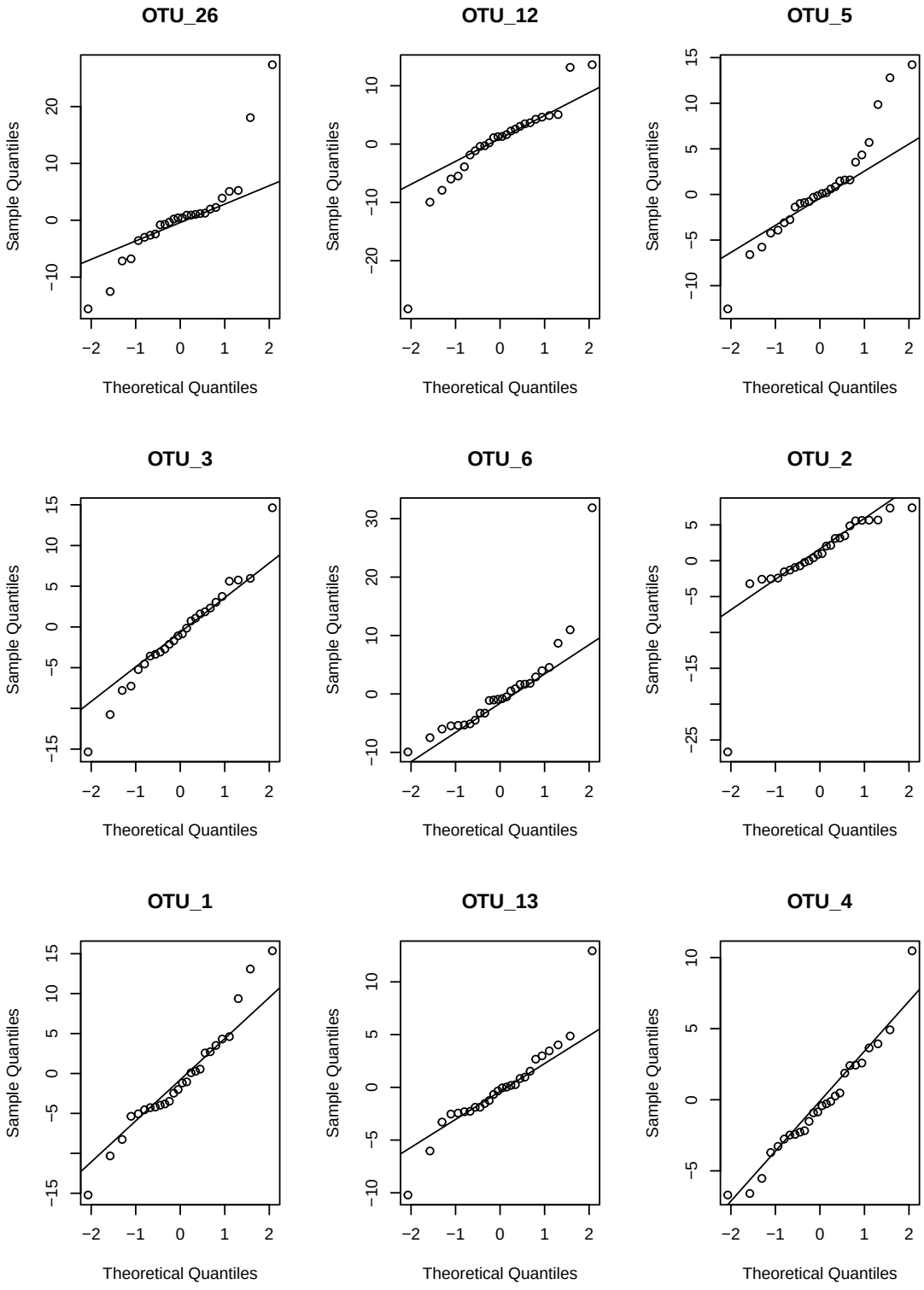
```
adf.test(res[-c(26),2])
```

```
  Augmented Dickey-Fuller Test

data:  res[-c(26), 2]
Dickey-Fuller = -2.4823, Lag order = 2, p-value = 0.3886
alternative hypothesis: stationary
```

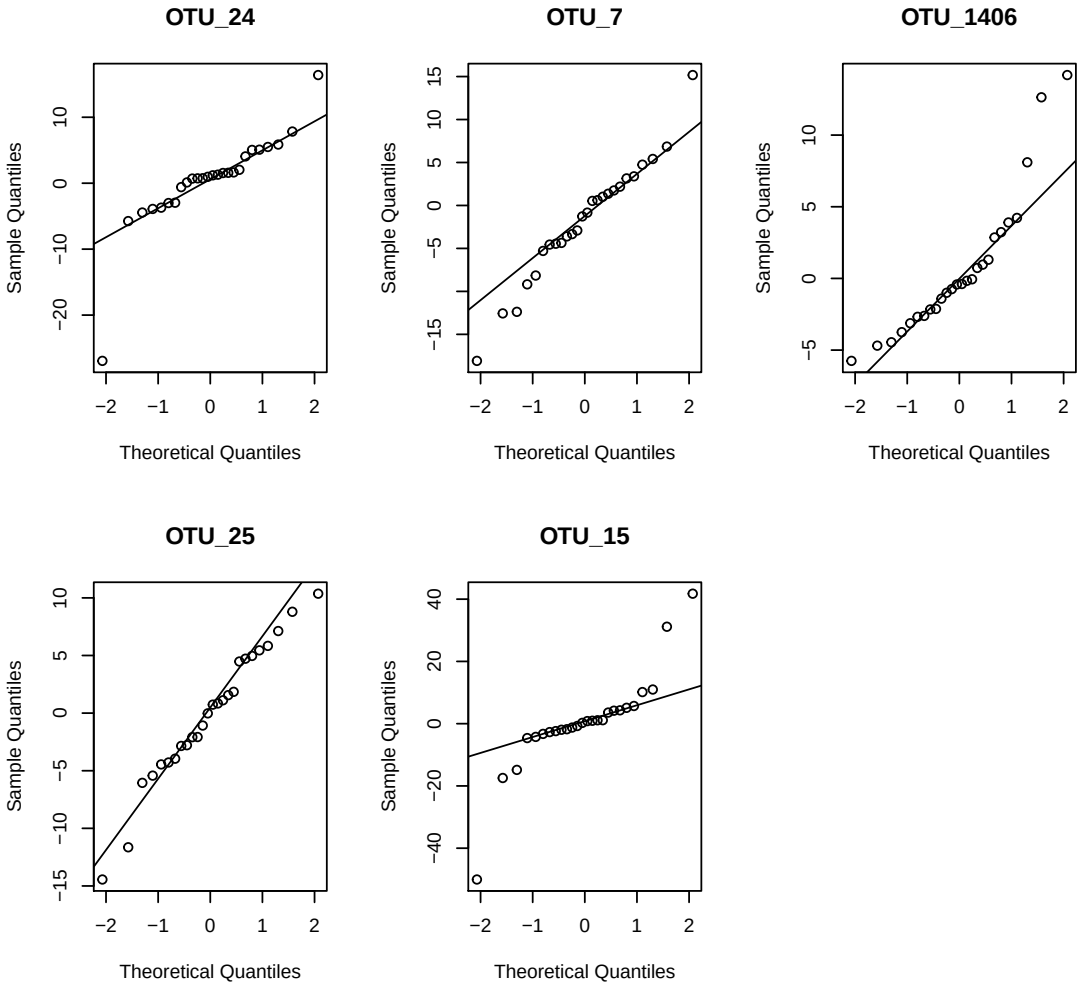
P-værdien falder fra $0,99$ til $0,38$ for Dickey-fuller testen når en observation fjernes. Dette er et kæmpe fald, hvilket tyder på, at der er for få observationer og at man derfor skal forholde sig kritisk overfor troværdigheden af Dickey-Fuller testen. Slutteligt undersøges om resdiualprocessen følger en normalfordeling. På figur 39 ses QQ-plots for residualerne, hvor det observeres, at residualerne approksimativt følger en normalfordeling med forekomst af enkle outliers.



Figur 38: Residualprocessen for de ikke-stationære tidsserier i VAR(1).



Figur 39: QQ plot af residualerne for de første 9 tidssserier i sparse VAR(1)



Figur 40: QQ plot af residualerne for de sidste 5 tidssserier i sparse VAR(1)

F Modellering af de enkelte OTU'er

I nuværende kapitel estimeres en model for hver af de 14 OTU'er i reaktor T14 ved hjælp af funktionen `modellering`. Dette skyldes både, fordi vi ønsker nogle bedre estimater for regressionskoefficienterne men også, fordi at ud fra VAR(1) modellen observeres, at de enkelte OTU'er korrelerer med flere af OTU'erne. Medtages alle disse parametre vil modellen ikke være repræsentativ i forhold til et lignende datasæt. Som nævnt i kapitel 2 fittes en model for de enkelte OTU'er baseret på den struktur, der er i VAR(1).

```
load("OTU_spline.Rda")
load("Akvivalent_tider.Rda")
data_samlet=as.data.frame(t(OTU_spline))
data_samlet[] <- lapply(data_samlet, as.character)
colnames(data_samlet) <- data_samlet[1, ]
data_samlet <- data_samlet[-1 ,]
T14_data=data_samlet[1:27,]
T14_data_matrix=data.matrix(T14_data, rownames.force = NA)
library(sparsevar)
set.seed(11) # 2 nuller,  lampda=0.2201965
T14_fit=fitVAR(T14_data_matrix, p=1)
T14_B=T14_fit$A[[1]]
T14_fit$lambda; diag(T14_fit$A[[1]])
```

```
[1] 0.2201965
 [1] 0.17764129 0.07183886 0.00000000 0.06781112 0.36354260 0.34330169 0.89960228
 [8] 0.55734649 0.53908805 0.49428956 0.34502379 0.00000000 0.08314770 0.81059709
```

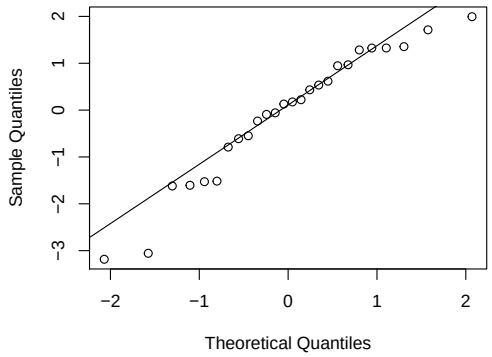
OTU 26

```
modellering(T14_data_matrix,T14_B,1)
```

```
[[1]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  data[c(NA, 1:26), omgang]
      0.5245    38.3895                     0.0284
s.e.  0.1856     0.8482                     0.0164

sigma^2 estimated as 1.592:  log likelihood = -43.09,  aic = 94.19

[[2]]
                      ar1                 intercept data[c(NA, 1:26), omgang]
              0.004701517               0.000000000               0.084281715

[[3]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept
      0.6480    39.7276
s.e.  0.1406     0.6845

sigma^2 estimated as 1.781:  log likelihood = -46.37,  aic = 98.75

[[4]]
         ar1    intercept
4.064072e-06 0.000000e+00
```

Modellen for OTU 26 er bestemt til $x_{1t} = 39,73 + 0,65x_{1t-1} + w_t$. Der undersøges for om residualerne opfylder antagelserne for om den er normalfordelt.

```
T14_OTU26=modellering(T14_data_matrix,T14_B,1)[length(modellering(T14_data_matrix,T14_B,1))-1]
T14_OTU26_res=T14_OTU26[[1]]$residuals[-1]
qqnorm(T14_OTU26_res,main=" "); qqline(T14_OTU26_res, main=" ")
```



Figur 41: QQ-plot over residualerne for AR(1) for abundansen af OTU 26.

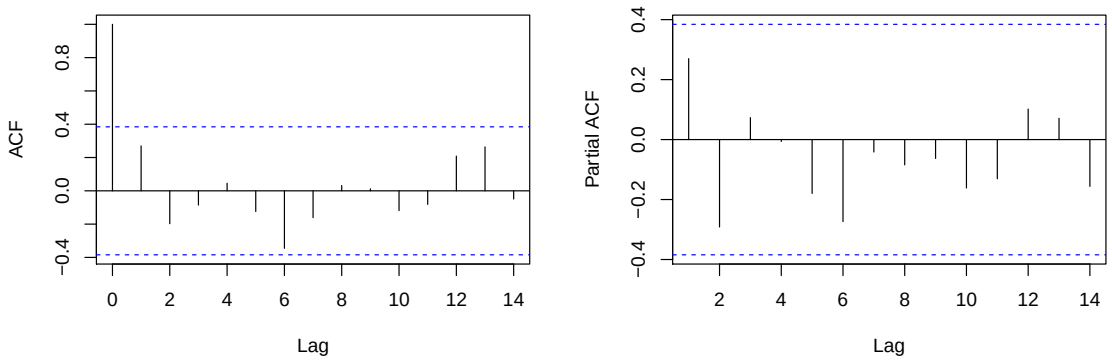
Ud fra QQ-plottet ses en tendens til at residualerne tilnærmelsesvis følger en normalfordeling. Der undersøges for om residualprocessen følger seriel korrelation.

```
Box.test(T14_OTU26_res)
```

```
  Box-Pierce test

data:  T14_OTU26_res
X-squared = 1.8933, df = 1, p-value = 0.1688
```

Nulhypotesen i Box-Pierce testen er, at ACF er lig nul for for alle lags. Testen viser, at man ikke kan forkaste nulhypotesen om, at der ingen seriel korrelation er. Der ses nu på korrelogrammet for hhv. PACF og ACF.

```
acf(T14_OTU26_res, main=" "); pacf(T14_OTU26_res, main=" ")
```



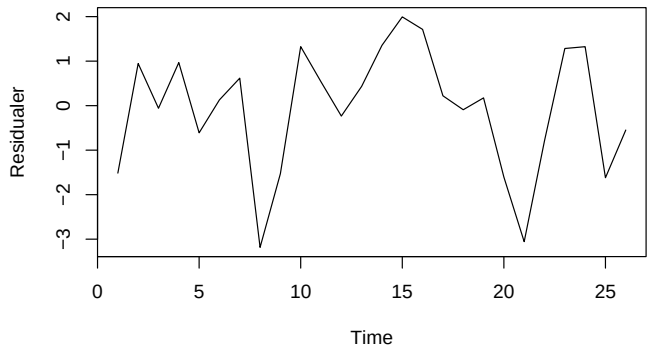
Figur 42: ACF- og PACF korrelogram af residualerne for abundansen af OTU 26.

Korrelogrammerne viser det samme som testen. Til sidst tjekkes om modellen opfylder stationaritet.

```
adf.test(T14_OTU26_res,k=0);adf.test(T14_data_matrix[,1],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU26_res
Dickey-Fuller = -3.6771, Lag order = 0, p-value = 0.0445
alternative hypothesis: stationary


  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 1]
Dickey-Fuller = -2.1598, Lag order = 0, p-value = 0.5115
alternative hypothesis: stationary
```

Testene viser, at residualprocessen er stationær, men observationerne for OTU26 er ikke stationær. Dermed er AR(1) ikke stationær.



Figur 43: Tidsserie af residualerne for OTU 26.

OTU 12

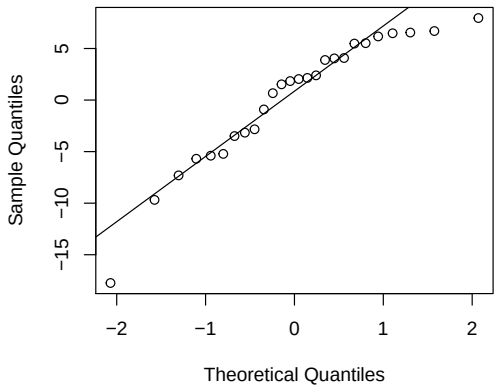
```
modellering(T14_data_matrix,T14_B,2)
```

```
[[1]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  OTU_26   OTU_1    OTU_4   OTU_24  OTU_15
      0.4084   -12.1117  2.6291  0.2101  -1.0251  -0.2231  0.1418
s.e.  0.2333    58.4774  1.2337  0.0843   0.4585   0.0767  0.1998

sigma^2 estimated as 27.61:  log likelihood = -80.12,  aic = 176.24

[[2]]
        ar1   intercept       OTU_26       OTU_1       OTU_4      OTU_24      OTU_15
0.079947389 0.835918171 0.033082932 0.012735315 0.025367171 0.003640102 0.477719485

[[3]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  OTU_26   OTU_1    OTU_4   OTU_24
      0.4205     9.8483  2.0481  0.1784  -0.8388  -0.2330
s.e.  0.2482    51.1373  0.9518  0.0729   0.3834   0.0762
```

```
sigma^2 estimated as 28.13:  log likelihood = -80.37,  aic = 174.74

[[4]]
         ar1   intercept       OTU_26       OTU_1       OTU_4      OTU_24
0.090233468 0.847283165 0.031417658 0.014373473 0.028662152 0.002225151

[[5]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
          ar1  intercept    OTU_1    OTU_4   OTU_24
      0.3765    80.1118  0.1352  -0.6034  -0.2363
s.e.  0.2948    40.1063  0.0775   0.4195   0.0938

sigma^2 estimated as 34.07:  log likelihood = -82.84,  aic = 177.68

[[6]]
       ar1  intercept       OTU_1       OTU_4      OTU_24
0.20152471 0.04577225 0.08132437 0.15032904 0.01176585

[[7]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
          ar1  intercept    OTU_1   OTU_24
      0.5374    26.7898  0.1764  -0.2218
s.e.  0.2206    13.9214  0.0761   0.0899

sigma^2 estimated as 36.22:  log likelihood = -83.73,  aic = 177.46

[[8]]
       ar1  intercept       OTU_1      OTU_24
0.01486655 0.05430910 0.02048264 0.01363696
```

Modellen for OTU 26 er bestemt til $x_{2t} = 26,79 + 0,54x_{2t-1} + 0,18x_{7t-1} - 0,22x_{10t-1} + w_t$.
Der undersøges for om residualerne er normalfordelte.

```
1 T14_OTU12=modellering(T14_data_matrix,T14_B,2)[length(modellering(T14_data_matrix,T14_B,2))-1]
2 T14_OTU12_res=T14_OTU12[[1]]$residuals[-1]
3 qqnorm(T14_OTU12_res,main=" "); qqline(T14_OTU12_res, main=" ")
```



Figur 44: QQ-plot over residualerne for for abundansen af OTU 12.

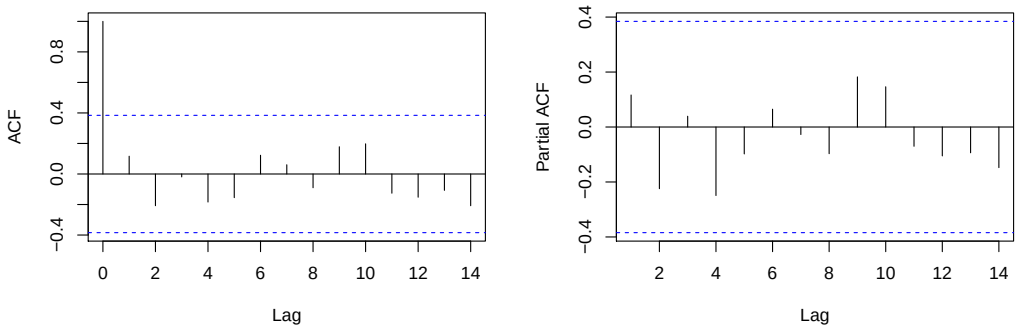
Ud fra QQ-plot ses en tendens til at residualerne tilnærmelsesvis følger en normalfordeling.
Der undersøges for om residualprocessen følger seriel korrelation.

```
1 Box.test(T14_OTU12_res)
```

```
  Box-Pierce test

data:  T14_OTU12_res
X-squared = 0.35281, df = 1, p-value = 0.5525
```

Box-Pierce testen viser, at nulhypotesen om, at der ingen seriel korrelation ikke kan forkastes. Korrelogrammet på figur 45 viser, at der ikke er en signifikant seriel korrelation.

```
1 acf(T14_OTU12_res, main=" "); pacf(T14_OTU12_res, main=" ")
```



Figur 45: ACF- og PACF korrelogram af residualerne for OTU 12.

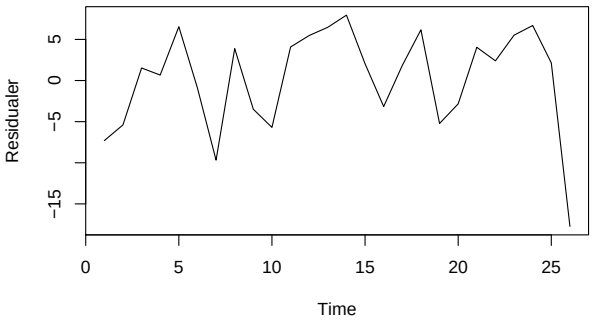
Til sidst tjekkes for stationaritet.

```
1 adf.test(T14_OTU12_res,k=0);adf.test(T14_data_matrix[,2],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU12_res
Dickey-Fuller = -2.9144, Lag order = 0, p-value = 0.224
alternative hypothesis: stationary


  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 2]
Dickey-Fuller = -1.1973, Lag order = 0, p-value = 0.8789
alternative hypothesis: stationary
```

Testene viser, at residualprocessen ikke er stationær. Vi ser nu på tidserien for residualerne.

```
plot.ts(T14_OTU12_res,ylab="Residualer")
```



Figur 46: Tidsserier af residualerne for AR(1) for abundansen af OTU 12.

Af figur 46 ser det ud til at både middelværdien og variansen er konstant op til tid 25. Der testes derfor om stationaritet er opfyldt når de sidste to observationer ikke er medtaget. Er dette tilfældet vil det indikere i, at der er for få observationer til at afgøre om residualprocessen opfylder stationaritet.

```
1 adf.test(T14_OTU12[[1]]$residuals[-c(1,26,27)],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU12[[1]]$residuals[-c(1, 26, 27)]
Dickey-Fuller = -4.0002, Lag order = 0, p-value = 0.02325
alternative hypothesis: stationary
```

OTU 5

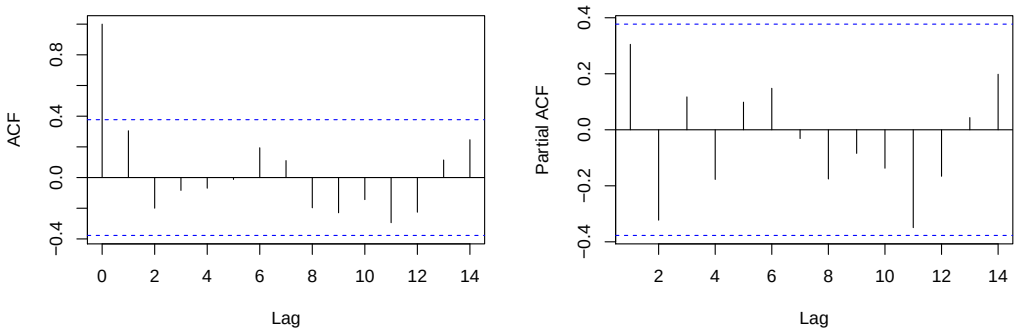
```
1 modellering(T14_data_matrix,T14_B,3)
```

```
Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_3   OTU_6   OTU_13   OTU_4    OTU_7
      0.0836   103.1044  -0.9920  0.3345  -0.3862  0.7321  -0.3507
s.e.  0.2230    24.2671   0.3464  0.2671   0.1108  0.2275   0.1991

sigma^2 estimated as 19.7:  log likelihood = -75.64,  aic = 167.28

[[2]]
         ar1    intercept         OTU_3         OTU_6        OTU_13         OTU_4
7.076936e-01 2.149876e-05 4.186371e-03 2.104921e-01 4.903053e-04 1.294116e-03
       OTU_7
7.814258e-02

[[3]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_3   OTU_13   OTU_4    OTU_7
      0.1038   113.9821  -0.9578  -0.3942  0.7728  -0.3654
s.e.  0.2252    23.5575   0.3549   0.1149  0.2372   0.2066

sigma^2 estimated as 20.89:  log likelihood = -76.41,  aic = 166.82

[[4]]
         ar1    intercept         OTU_3        OTU_13         OTU_4         OTU_7
6.448447e-01 1.308506e-06 6.958083e-03 5.997276e-04 1.121835e-03 7.704692e-02

[[5]]

Call:
```

```
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_3   OTU_13   OTU_4
      0.0995   109.8732  -1.0278  -0.4679  0.7237
s.e.  0.2321    24.7905   0.3756   0.1139  0.2496

sigma^2 estimated as 23.49:  log likelihood = -77.93,  aic = 167.87

[[6]]
         ar1    intercept        OTU_3       OTU_13        OTU_4
6.680702e-01 9.333516e-06 6.208590e-03 3.966701e-05 3.737726e-03

[[7]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept   OTU_13   OTU_4
      0.1265    60.9576  -0.3612  0.4121
s.e.  0.2210    20.1985   0.1229  0.2526

sigma^2 estimated as 30.4:  log likelihood = -81.29,  aic = 172.57

[[8]]
        ar1   intercept      OTU_13       OTU_4
0.566969813 0.002545087 0.003298044 0.102761403

[[9]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  data[c(NA, 1:26), omgang]
      0.2428    93.1783                    -0.3823
s.e.  0.2287     6.1793                     0.1435

sigma^2 estimated as 33.14:  log likelihood = -82.43,  aic = 172.86

[[10]]
                     ar1                intercept data[c(NA, 1:26), omgang]
            2.884053e-01             2.220321e-51             7.713677e-03

[[11]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept
      0.3506    77.0302
s.e.  0.1932     1.8574

sigma^2 estimated as 40.42:  log likelihood = -88.32,  aic = 182.63

[[12]]
       ar1  intercept
0.06962561 0.00000000
```

Idet regressionskoefficienten for lag en ikke er signifikant forskellig fra nul, følger abundansen af OTU 5 ikke en AR(1), hvilket også støttes op af ACF og PACF korrelogrammerne.

```
1 acf(T14_data_matrix[,3], main=" "); pacf(T14_data_matrix[,3], main=" ")
```



Figur 47: ACF- og PACF korrelogram af abundansen for OTU 5.

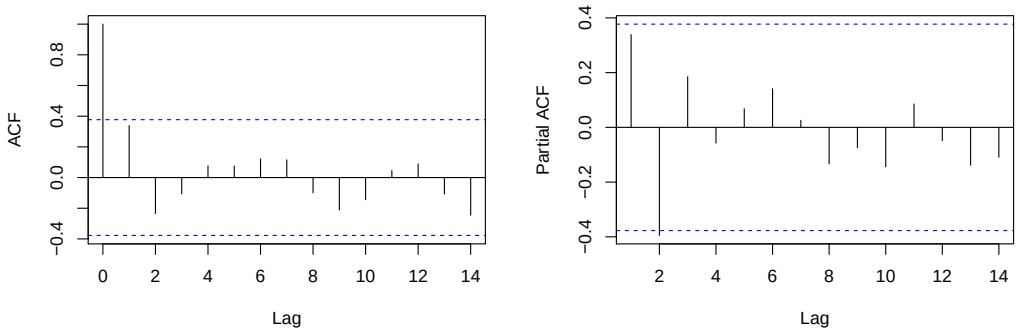
OTU 3

```
1 modellering(T14_data_matrix,T14_B,4)
```

```
[[1]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  data[c(NA, 1:26), omgang]
      0.1743    70.4666                   -0.1044
s.e.  0.1989     2.5819                    0.0598

sigma^2 estimated as 6.609:  log likelihood = -61.46,  aic = 130.91

[[2]]
                     ar1                intercept data[c(NA, 1:26), omgang]
             3.808749e-01             5.253117e-164              8.084309e-02

[[3]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept
      0.4058    66.4949
s.e.  0.1930     0.9146

sigma^2 estimated as 8.164:  log likelihood = -66.75,  aic = 139.49

[[4]]
       ar1  intercept
0.03547434 0.00000000
```

Abundansen af OTU 3 kan beskrives som $x_{4t} = 66,49 + 0,41x_{4t-1} + w_t$. ACF og PACF korrellogramerne for OTU 3 ses på figur 48.

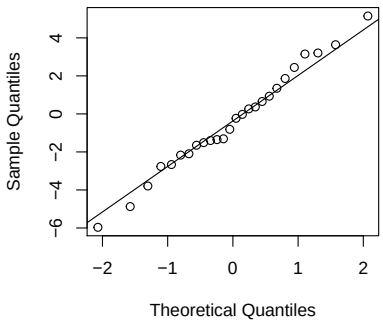
```
1 acf(T14_data_matrix[,4], main=" "); pacf(T14_data_matrix[,4], main=" ")
```



Figur 48: Korrelogram af ACF og PACF for abundansen af OTU 3.

Der undersøges for om residualerne er normalfordelte.

```
T14_OTU3=modellering(T14_data_matrix,T14_B,4)[length(modellering(T14_data_matrix,T14_B,4))-1]
T14_OTU3_res=T14_OTU3[[1]]$residuals[-1]
qqnorm(T14_OTU3_res,main=" "); qqline(T14_OTU3_res, main=" ")
```



Figur 49: QQ-plot af residualerne for OTU 3.

Af QQ-plottet følger residualerne tilnærmelsesvis en normalfordeling.

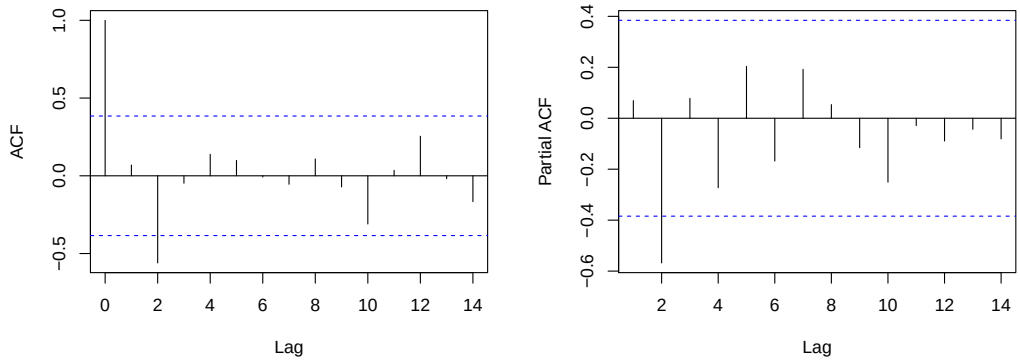
```
Box.test(T14_OTU3_res)
```

```
  Box-Pierce test

data:  T14_OTU3_res
X-squared = 0.12491, df = 1, p-value = 0.7238
```

Af Box-Pierce testen kan man ikke forkaste, at der ingen seriel korrelation er, dog viser korrelogrammerne, se figur 50, at der er en signifikant korrelation for lag 2. Ved at sammenholde figur 48 og figur 50 er der en tendens til at AR(2) vil være en bedre model.

```
  Box-Pierce test

data:  T14_OTU3_res
X-squared = 0.12491, df = 1, p-value = 0.7238
```

```
acf(T14_OTU3_res, main=" "); pacf(T14_OTU3_res, main=" ")
```



Figur 50: ACF- og PACF korrelogram af residualerne for OTU 3.

```
1 adf.test(T14_OTU3_res,k=0);adf.test(T14_data_matrix[,4],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU3_res
Dickey-Fuller = -4.4214, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(T14_OTU3_res, k = 0) : p-value smaller than printed p-value

  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 4]
Dickey-Fuller = -3.6883, Lag order = 0, p-value = 0.04337
alternative hypothesis: stationary
```

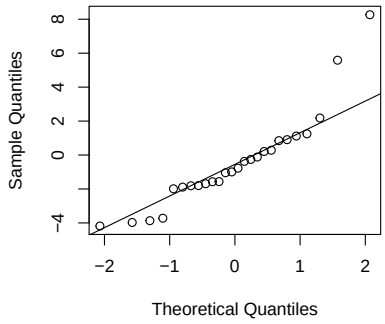
Tidsserien er altså stationær.

OTU 6

```
1 modellering(T14_data_matrix,T14_B,5)
```

```
[[1]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  data[c(NA, 1:26), omgang]
      0.6219    36.5991                    0.1755
s.e.  0.1585     9.1042                    0.1669

sigma^2 estimated as 6.693:  log likelihood = -61.85,  aic = 131.7

[[2]]
                     ar1                intercept data[c(NA, 1:26), omgang]
             8.679491e-05             5.819233e-05              2.931917e-01

[[3]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept
      0.6851    46.7703
s.e.  0.1584     1.7076

sigma^2 estimated as 8.081:  log likelihood = -66.84,  aic = 139.67

[[4]]
          ar1     intercept
 1.519791e-05 3.601724e-165
```

Modellen for OTU 6 er bestemt til $x_{5t} = 46,77 + 0.69x_{5t-1} + w_t$. Korrelogrammerne for ACF og PACF ses på figur



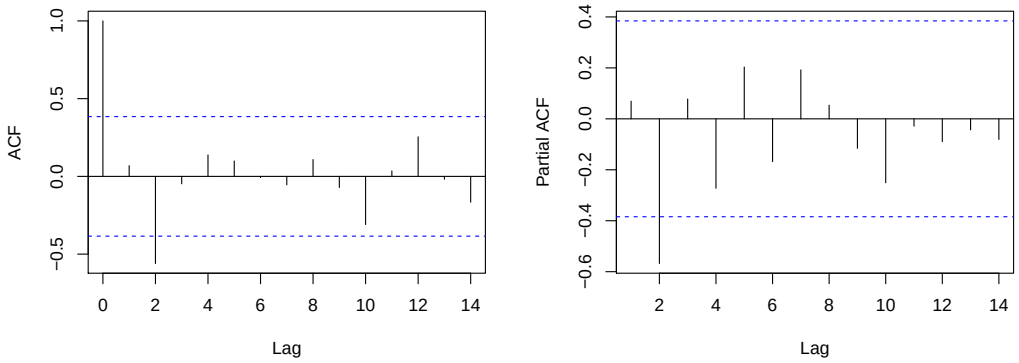
Figur 51: QQ-plot af residualerne for OTU 6.

```
1 Box.test(T14_OTU6_res)
```

```
  Box-Pierce test

data:  T14_OTU6_res
X-squared = 1.5027, df = 1, p-value = 0.2203
```

Af Box-Pierce testen kan man ikke forkaste, at der ingen seriel korrelation er. Korrelogrammerne på figur 52 viser, at der ikke er en signifikant korrelation.

```
1 acf(T14_OTU6_res, main=" "); pacf(T14_OTU6_res, main=" ")
```



Figur 52: ACF- og PACF korrelogram af residualerne for OTU 6.

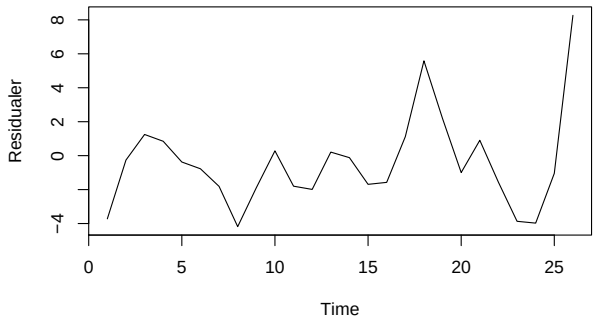
```
1 adf.test(T14_OTU6_res,k=0);adf.test(T14_data_matrix[,5],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU6_res
Dickey-Fuller = -2.3352, Lag order = 0, p-value = 0.4447
alternative hypothesis: stationary
```

```
  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 5]
Dickey-Fuller = -2.3384, Lag order = 0, p-value = 0.4433
alternative hypothesis: stationary
```

Dickey-Fuller testene viser, at tidsserien ikke er stationær.

```
1 plot.ts(T14_OTU6_res,ylab="Residualer")
```



Figur 53: Tidsserie af residualerne for OTU 6.

OTU 2

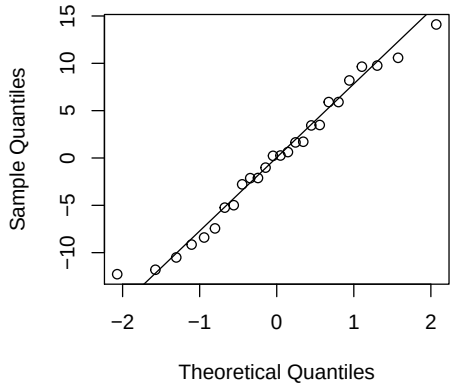
```
1 modellering(T14_data_matrix,T14_B,6)
```

```
[[1]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
          ar1  intercept   OTU_26   OTU_5    OTU_6   OTU_13  OTU_24
      -0.0393   183.9069  -0.5317  0.5426  -0.4960  -0.3176  0.2875
s.e.   0.2411    38.6753   0.7295  0.2746   0.3705   0.1727  0.0589

sigma^2 estimated as 37.12:  log likelihood = -83.88,  aic = 183.76

[[2]]
         ar1    intercept       OTU_26        OTU_5        OTU_6       OTU_13       OTU_24
8.706416e-01 1.982976e-06 4.660687e-01 4.813965e-02 1.806649e-01 6.596065e-02 1.070786e-06

[[3]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
          ar1  intercept   OTU_5    OTU_6   OTU_13  OTU_24
      -0.0402   165.1470  0.5415  -0.5492  -0.2957  0.2688
s.e.   0.2417    28.8129  0.2793   0.3635   0.1715  0.0537

sigma^2 estimated as 37.89:  log likelihood = -84.14,  aic = 182.29

[[4]]
         ar1    intercept        OTU_5        OTU_6       OTU_13       OTU_24
8.679621e-01 9.942788e-09 5.250743e-02 1.308469e-01 8.458616e-02 5.652182e-07

[[5]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept   OTU_5   OTU_13  OTU_24
      0.1580   150.4571  0.4157  -0.3159  0.2769
```

```
s.e.  0.2258    28.9300  0.2981   0.2024  0.0642

sigma^2 estimated as 40.44:  log likelihood = -85,  aic = 182.01

[[6]]
         ar1    intercept        OTU_5       OTU_13       OTU_24
4.840179e-01 1.985136e-07 1.632348e-01 1.185840e-01 1.601497e-05

[[7]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept   OTU_13  OTU_24
      0.2839   190.0669  -0.4805  0.2715
s.e.  0.1975     7.2423   0.1885  0.0726

sigma^2 estimated as 43.31:  log likelihood = -85.92,  aic = 181.85

[[8]]
         ar1     intercept        OTU_13        OTU_24
 1.507107e-01 8.401492e-152  1.078984e-02  1.841906e-04

[[9]]

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  data[c(NA, 1:26), omgang]
      0.4788   173.9186                     0.1850
s.e.  0.1777     4.4004                     0.0808

sigma^2 estimated as 51.78:  log likelihood = -88.33,  aic = 184.67

[[10]]
                     ar1                intercept data[c(NA, 1:26), omgang]
             0.007050494             0.000000000               0.022070791
```

Abundansen af OTU 2 kan beskrives ved $x_{6t} = 173,92 + 0,48x_{6t-1} + 0.19x_{10t-1} + w_t$. Af figur 59 følger residualerne tilnærmelsesvis en normalfordeling.

```
T14_OTU2=modellering(T14_data_matrix,T14_B,6)[length(modellering(T14_data_matrix,T14_B,6))-1]
T14_OTU2_res=T14_OTU2[[1]]$residuals[-1]
par(mfrow=c(1,1))
qqnorm(T14_OTU2_res,main=" "); qqline(T14_OTU2_res, main=" ")
```



Figur 54: QQ-plot af residualerne for OTU 2.

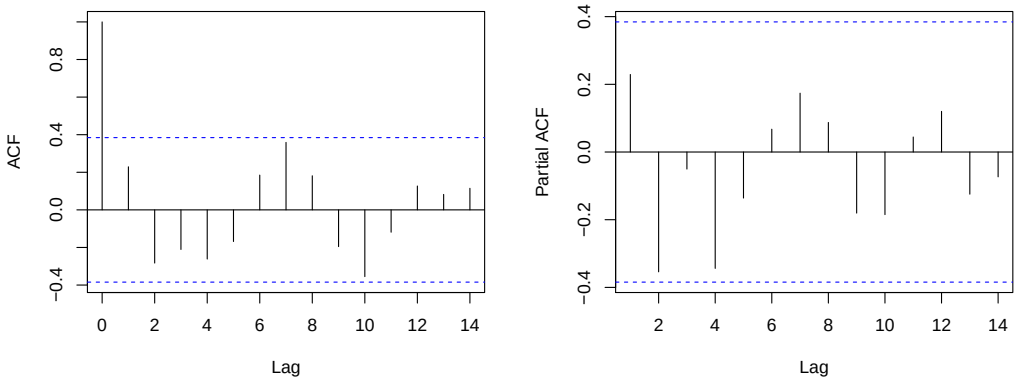
```
Box.test(T14_OTU2_res)
```

```
  Box-Pierce test
```

```
data:  T14_OTU2_res
X-squared = 1.3649, df = 1, p-value = 0.2427
```

Af Box-Pierce testen kan man ikke forkaste, at der ingen seriel korrelation er. ACF og PACF korrelogrammerne for residualerne ses på figur 56.

```
1 acf(T14_OTU2_res, main=" "); pacf(T14_OTU2_res, main=" ")
```



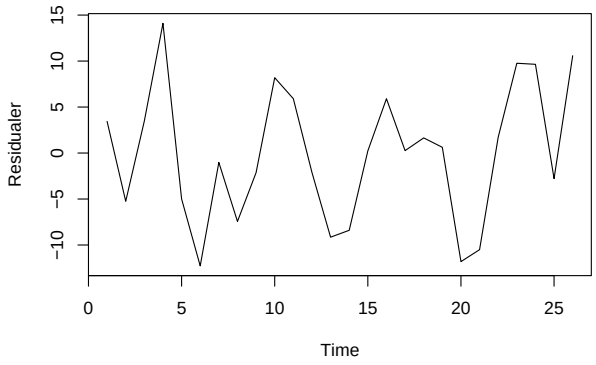
Figur 55: ACF og PACF korrelogrammer for abundansen af OTU 2.

```
1 adf.test(T14_OTU2_res,k=0);adf.test(T14_data_matrix[,6],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU2_res
Dickey-Fuller = -3.5509, Lag order = 0, p-value = 0.05682
alternative hypothesis: stationary


  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 6]
Dickey-Fuller = -2.6048, Lag order = 0, p-value = 0.3416
alternative hypothesis: stationary
```

Dickey-Fuller testen viser, at tidsserien ikke er stationær.

```
1 plot.ts(T14_OTU2_res,ylab="Residualer")
```



Figur 56: Tidsserie af residualerne for OTU 2.

```
adf.test(T14_OTU2[[1]]$residuals[-c(1,4)],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU2[[1]]$residuals[-c(1, 4)]
Dickey-Fuller = -3.8548, Lag order = 0, p-value = 0.0318
alternative hypothesis: stationary
```

Ved at fjerne en af observationerne, viser Dickey-Fuller testen at være stationær. Dette tyder altså på, at der er for få observationer, til at konkludere om den opfylder stationaritet, og dermed er Dickey-Fuller testen ikke valid.

OTU 1

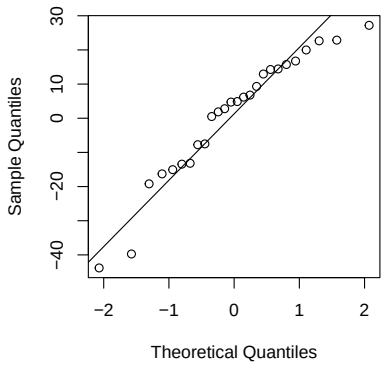
```
modellering(T14_data_matrix,T14_B,7)
```

```
         ar1   intercept      OTU_26      OTU_12       OTU_6       OTU_2      OTU_13
0.560351120 0.986185635 0.001912241 0.795920447 0.315954547 0.085608429 0.485483395
     OTU_24       OTU_7
0.016114910 0.055647688

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  OTU_26  OTU_12   OTU_6    OTU_2  OTU_13   OTU_24    OTU_7
      0.2761     4.2751  8.9855  0.2223  2.3326  -1.2205  0.5453  -0.8263  -2.2112
s.e.  0.4741   246.9098  2.8953  0.8594  2.3261   0.7100  0.7818   0.3434   1.1554

sigma^2 estimated as 266.4:  log likelihood = -109.53,  aic = 239.07
         ar1    intercept       OTU_26        OTU_6        OTU_2       OTU_13
0.2450643996 0.5799536112 0.0044306381 0.2353984554 0.0607967198 0.5225907988
      OTU_24       OTU_7
0.0005595435 0.0625277246

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  OTU_26   OTU_6    OTU_2  OTU_13   OTU_24    OTU_7
      0.3681    63.3432  8.7042  1.8569  -1.2841  0.4288  -0.8868  -2.1556
s.e.  0.3166   114.4511  3.0587  1.5649   0.6849  0.6706   0.2570   1.1573

sigma^2 estimated as 266.3:  log likelihood = -109.56,  aic = 237.13
         ar1    intercept       OTU_26        OTU_6        OTU_2       OTU_24
0.2412974882 0.3692480644 0.0088959125 0.2099085655 0.0342530725 0.0002488752
       OTU_7
0.0699912202

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])
```

```
Coefficients:
         ar1  intercept  OTU_26   OTU_6    OTU_2   OTU_24    OTU_7
      0.3988    95.6895  8.3361  1.9992  -1.4469  -0.7919  -1.7078
s.e.  0.3403   106.5724  3.1865  1.5945   0.6834   0.2162   0.9425

sigma^2 estimated as 270.1:  log likelihood = -109.76,  aic = 235.53
         ar1     intercept        OTU_26        OTU_2       OTU_24        OTU_7
2.120172e-03 7.436465e-02 4.432754e-02 1.048761e-01 3.218886e-05 4.329891e-02

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  OTU_26    OTU_2   OTU_24    OTU_7
      0.6457   179.7312  7.0675  -1.0504  -0.9181  -1.9808
s.e.  0.2101   100.7257  3.5144   0.6477   0.2208   0.9802

sigma^2 estimated as 279:  log likelihood = -110.37,  aic = 234.73
         ar1     intercept        OTU_26       OTU_24        OTU_7
1.799886e-06 1.292863e-01 2.030174e-01 3.498173e-04 4.333682e-03

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  OTU_26   OTU_24    OTU_7
      0.7610   156.6626  3.3707  -0.8507  -2.6493
s.e.  0.1594   103.2766  2.6479   0.2379   0.9287

sigma^2 estimated as 302.6:  log likelihood = -111.59,  aic = 235.17
         ar1     intercept       OTU_24        OTU_7
9.927047e-08 6.042538e-09 5.839375e-04 1.138608e-02

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept   OTU_24    OTU_7
      0.7809   275.3162  -0.8455  -2.3142
s.e.  0.1466    47.3412   0.2458   0.9145

sigma^2 estimated as 320.8:  log likelihood = -112.38,  aic = 234.77
```

Hermed kan abundansen af OTU 1 beskrives ved $x_{7t} = 275,32 + 0,78x_{7t-1} - 0,85x_{10t-1} - 2,31x_{11t-1} + w_t$. Af figur 59 ses det, at residualerne tilnærmelsesvis følger en normalfordeling.

```
1 T14_OTU1=modellering(T14_data_matrix,T14_B,7)[length(modellering(T14_data_matrix,T14_B,7))-1]
2 T14_OTU1_res=T14_OTU1[[1]]$residuals[-1]
3 qqnorm(T14_OTU1_res,main=" "); qqline(T14_OTU1_res, main=" ")
```



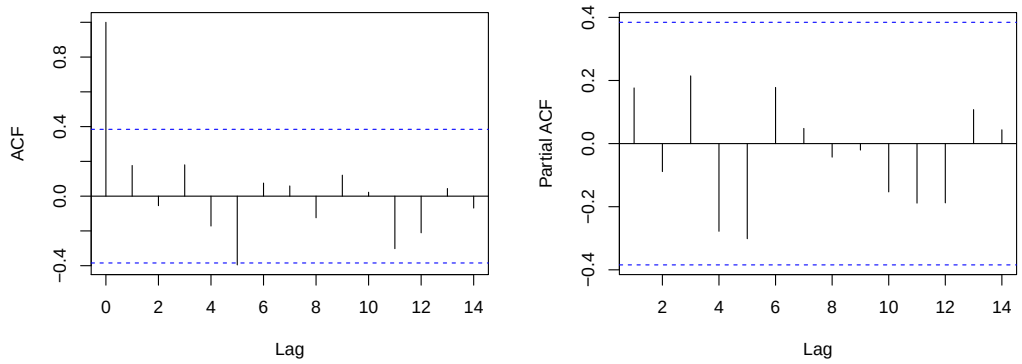
Figur 57: QQ-plot af residualerne for OTU 1.

```
1 Box.test(T14_OTU1_res)
```

```
  Box-Pierce test

data:  T14_OTU1_res
X-squared = 0.80876, df = 1, p-value = 0.3685
```

Af Box-Pierce testen kan man ikke forkaste, at der ingen seriel korrelation er. ACF korrelogrammet viser dog, at der er en signifikant korrelation for lag 5.

```
1 acf(T14_OTU1_res, main=" "); pacf(T14_OTU1_res, main=" ")
```

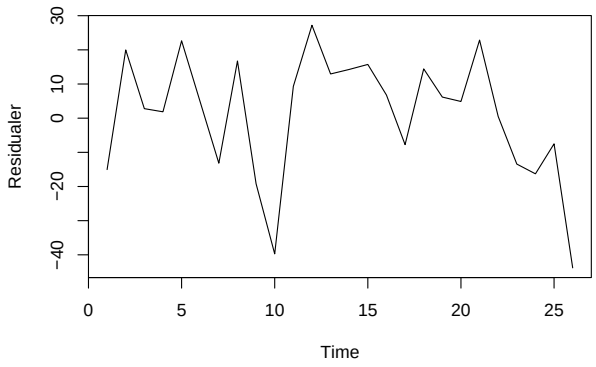


Figur 58: ACF og PACF korrelogrammer af residualerne for OTU 1.

```
1 adf.test(T14_OTU1_res,k=0);adf.test(T14_data_matrix[,7],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU1_res
Dickey-Fuller = -3.4819, Lag order = 0, p-value = 0.0664
alternative hypothesis: stationary


  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 7]
Dickey-Fuller = -0.32968, Lag order = 0, p-value = 0.9824
alternative hypothesis: stationary
```

Dickey-Fuller testene viser, at tidsserien ikke er stationær.



Figur 59: Tidsserie af residualerne for OTU 1.

```
adf.test(T14_OTU1[[1]]$residuals[-c(1,10)],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU1[[1]]$residuals[-c(1, 10)]
Dickey-Fuller = -4.256, Lag order = 0, p-value = 0.01433
alternative hypothesis: stationary
```

Ved at fjerne en af observationerne viser Dickey-Fuller testen, at residualprocessen er stationær.

OTU 13

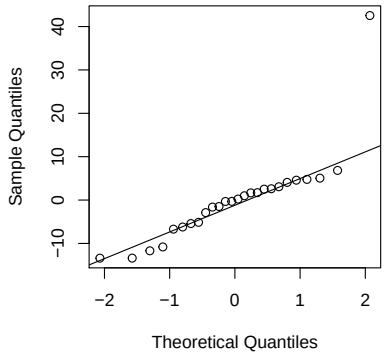
```
modellering(T14_data_matrix,T14_B,8)
```

```
       ar1  intercept     OTU_26      OTU_3      OTU_6      OTU_1      OTU_7   OTU_1406
0.47235551 0.13377714 0.11784740 0.79426375 0.55543385 0.15325511 0.02349514 0.57554284
    OTU_15
0.80282893

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_26    OTU_3    OTU_6    OTU_1   OTU_7  OTU_1406  OTU_15
      0.5132   153.4424  -2.8350  -0.2668  -0.6346  -0.2001  1.1248    0.4841  0.1404
s.e.  0.7142   102.3378   1.8128   1.0230   1.0763   0.1401  0.4965    0.8646  0.5624

sigma^2 estimated as 77.52:  log likelihood = -93.6,  aic = 207.2
       ar1  intercept     OTU_26      OTU_3      OTU_6      OTU_1      OTU_7   OTU_1406
0.14910391 0.04376474 0.05657622 0.85359048 0.28873185 0.05129390 0.01975992 0.48344520

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_26    OTU_3    OTU_6    OTU_1   OTU_7  OTU_1406
      0.4370   171.1615  -3.0916  -0.1360  -0.7499  -0.2143  1.1221    0.4622
s.e.  0.3029    84.8873   1.6215   0.7367   0.7068   0.1099  0.4814    0.6595

sigma^2 estimated as 78.21:  log likelihood = -93.67,  aic = 205.34
       ar1  intercept     OTU_26      OTU_6      OTU_1      OTU_7   OTU_1406
0.14300399 0.03755013 0.04620761 0.24565227 0.04430924 0.01746019 0.47778443

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_26    OTU_6    OTU_1   OTU_7  OTU_1406
      0.4327   165.5045  -3.1706  -0.7846  -0.2077  1.1331    0.4599
s.e.  0.2954    79.5797   1.5905   0.6758   0.1033  0.4767    0.6479

sigma^2 estimated as 78.33:  log likelihood = -93.69,  aic = 203.38
        ar1   intercept      OTU_26       OTU_6       OTU_1       OTU_7
0.062919068 0.019495138 0.054016734 0.307614171 0.086691305 0.008780291

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_26    OTU_6    OTU_1   OTU_7
      0.5449   191.9520  -3.2630  -0.7459  -0.2270  1.2143
s.e.  0.2930    82.1738   1.6936   0.7311   0.1325  0.4634

sigma^2 estimated as 79.3:  log likelihood = -93.92,  aic = 201.84
        ar1   intercept      OTU_26       OTU_1       OTU_7
0.007320906 0.022804663 0.021664649 0.051954539 0.013169470

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_26    OTU_1   OTU_7
      0.6548   181.0652  -3.6642  -0.2595  1.1699
s.e.  0.2442    79.5300   1.5958   0.1335  0.4719

sigma^2 estimated as 81.54:  log likelihood = -94.39,  aic = 200.77
        ar1   intercept      OTU_26       OTU_7
0.021168826 0.275802568 0.167320029 0.002390836
```

```
Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept   OTU_26   OTU_7
      0.5342    62.7607  -2.0428  1.5054
s.e.  0.2318    57.5894   1.4794  0.4957

sigma^2 estimated as 97.87:  log likelihood = -96.65,  aic = 203.3
                      ar1                intercept data[c(NA, 1:26), omgang]
              0.034462209              0.621974577              0.007754545

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  data[c(NA, 1:26), omgang]
      0.5177   -10.7751                     1.3219
s.e.  0.2448    21.8537                     0.4965

sigma^2 estimated as 105.2:  log likelihood = -97.58,  aic = 203.15
```

Abundansen af OTU 13 kan beskrives ved $x_{8t} = -10,78 + 0,52x_{8t-1} + 1,32x_{11t-1} + w_t$. Af figur 60 ses det, at residualerne tilnærmelsesvis følger en normalfordeling. Det bemærkes også at en af residualerne opfører sig underligt.

```
1 T14_OTU13_res=T14_OTU13[[1]]$residuals[-1]
2 qqnorm(T14_OTU13_res,main=" "); qqline(T14_OTU13_res, main=" ")
```



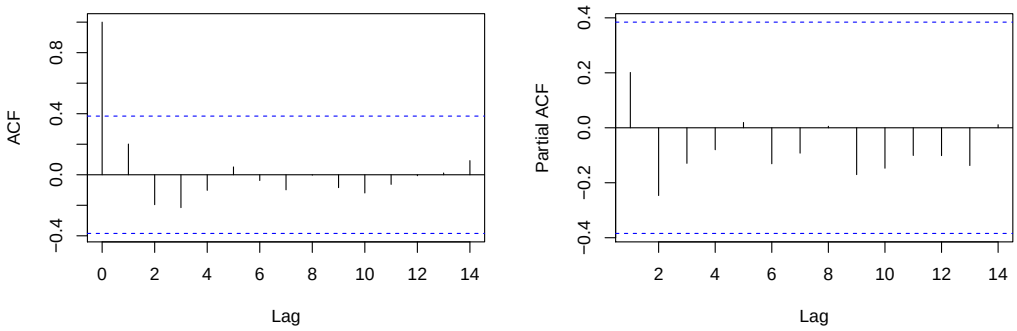
Figur 60: QQ-plot af residualerne for OTU 1.

```
1 Box.test(T14_OTU13_res)
```

```
  Box-Pierce test

data:  T14_OTU13_res
X-squared = 1.0529, df = 1, p-value = 0.3048
```

Af Box-Pierce testen kan man ikke forkaste, at der ingen seriel korrelation er. Dette passer godt med ACF korrelogrammet, se figur 61.

```
1 acf(T14_OTU13_res, main=" "); pacf(T14_OTU13_res, main=" ")
```



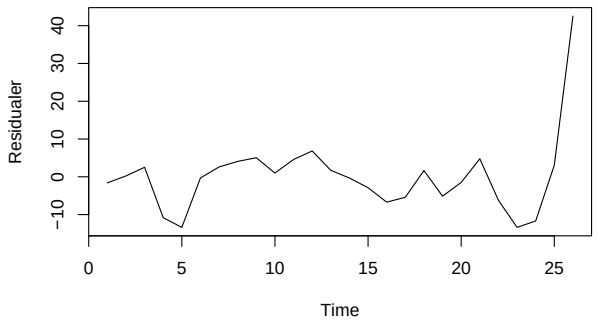
Figur 61: ACF og PACF korrelogrammer for OTU 13.

```
adf.test(T14_OTU13_res,k=0);adf.test(T14_data_matrix[,8],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU13_res
Dickey-Fuller = -0.7731, Lag order = 0, p-value = 0.9522
alternative hypothesis: stationary


  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 8]
Dickey-Fuller = -0.9004, Lag order = 0, p-value = 0.9356
alternative hypothesis: stationary
```

Dickey-fuller testene viser, at tidsserien ikke er stationær.

```
plot.ts(T14_OTU13_res,ylab="Residualer")
```



Figur 62: Tidsserie af residualerne for OTU 13.

Ved at fjerne to af observationerne fader p-værdien markant for Dickey-Fuller testen, hvilket igen viser, at der er for få observationer til at afgøre stationaritet med Dickey-Fuller testen.

```
adf.test(T14_OTU13[[1]]$residuals[-c(1,25,27)],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU13[[1]]$residuals[-c(1, 25, 27)]
Dickey-Fuller = -3.2625, Lag order = 0, p-value = 0.09688
alternative hypothesis: stationary
```

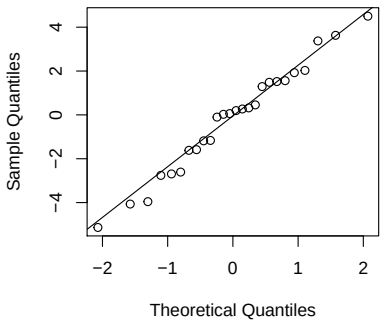
OTU 4

```
modellering(T14_data_matrix,T14_B,9)
```

```
         ar1    intercept       OTU_24     OTU_1406
1.919269e-12 4.792224e-18 1.310058e-02 6.776097e-01

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
          ar1  intercept  OTU_24  OTU_1406
      0.8041    75.3899  0.0756   -0.0623
s.e.  0.1142     8.7073  0.0305    0.1499

sigma^2 estimated as 5.731:  log likelihood = -60.11,  aic = 130.22
                      ar1                 intercept data[c(NA, 1:26), omgang]
             3.920616e-13             1.469734e-176              9.700251e-03

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
          ar1  intercept  data[c(NA, 1:26), omgang]
      0.8109    71.9195                     0.0778
s.e.  0.1117     2.5386                     0.0301

sigma^2 estimated as 5.762:  log likelihood = -60.2,  aic = 128.39
```

Abundansen af OTU 4 kan beskrives ved $x_{9t} = 71,92 + 0,82x_{9t-1} + 0,08x_{10t-1} + w_t$. Af figur 63 ses det, at residualerne tilnærmelsesvis følger en normalfordeling.

```
T14_OTU4=modellering(T14_data_matrix,T14_B,9)[length(modellering(T14_data_matrix,T14_B,9))-1]
T14_OTU4_res=T14_OTU4[[1]]$residuals[-1]
qqnorm(T14_OTU4_res,main=" "); qqline(T14_OTU4_res, main=" ")
```



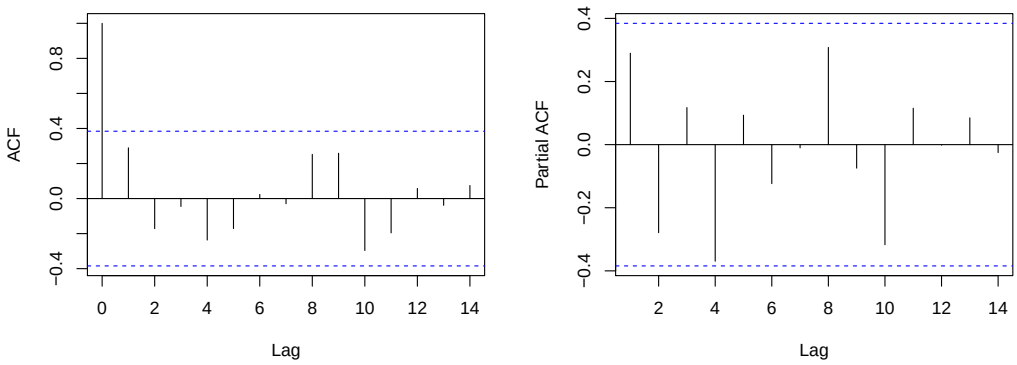
Figur 63: QQ-plot af residualerne for OTU 4.

```
Box.test(T14_OTU4_res)
```

```
  Box-Pierce test

data:  T14_OTU4_res
X-squared = 2.1822, df = 1, p-value = 0.1396
```

Af Box-Pierce testen kan man ikke forkaste, at der ingen seriel korrelation er. Dette understøttes af ACF korrelogrammet på figur 64.

```
acf(T14_OTU4_res, main=" "); pacf(T14_OTU4_res, main=" ")
```



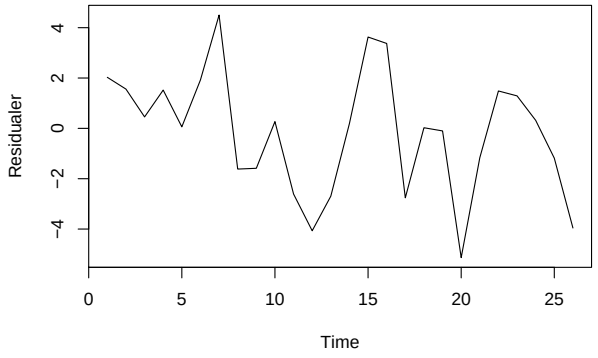
Figur 64: ACF og PACF korrelogrammer for OTU 13.

```
adf.test(T14_OTU4_res,k=0);adf.test(T14_data_matrix[,9],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU4_res
Dickey-Fuller = -3.4654, Lag order = 0, p-value = 0.0687
alternative hypothesis: stationary


  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 9]
Dickey-Fuller = -3.0372, Lag order = 0, p-value = 0.1765
alternative hypothesis: stationary
```

Dickey-Fuller testene viser, at tidsserien ikke er stationær.

```
plot.ts(T14_OTU4_res,ylab="Residualer")
```



Figur 65: Tidsserie af residualerne for OTU 4.

Ved at fjerne en af observationerne bliver residualprocessen signifikant stationær.

```
adf.test(T14_OTU4[[1]]$residuals[-c(1,15)],k=0)
```

```
  Augmented Dickey-Fuller Test
```

```
data:  T14_OTU4[[1]]$residuals[-c(1, 15)]
Dickey-Fuller = -3.7202, Lag order = 0, p-value = 0.04142
alternative hypothesis: stationary
```

OTU 24

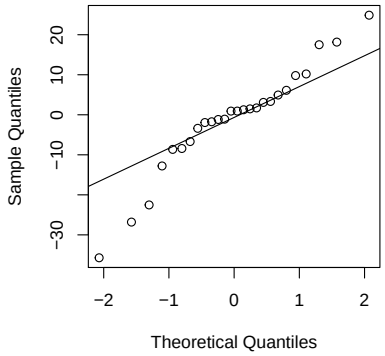
```
1 modellering(T14_data_matrix,T14_B,12)
```

```
         ar1    intercept       OTU_26        OTU_5        OTU_3        OTU_6
0.0126052863 0.0008054477 0.7428132173 0.0050724060 0.0586153070 0.4217056869
       OTU_2        OTU_1        OTU_4       OTU_25
0.6074777617 0.0030760111 0.2921215777 0.0230238573

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_26    OTU_5   OTU_3   OTU_6   OTU_2    OTU_1    OTU_4
      0.6037   382.7449   -0.8703  -1.2698  3.0987  1.4760  0.2576  -0.4884  -1.2013
s.e.  0.2420   114.2210    2.6523   0.4531  1.6386  1.8371  0.5015   0.1650   1.1403
       OTU_25
      -4.4918
s.e.   1.9761

sigma^2 estimated as 147.5:  log likelihood = -102.03,  aic = 226.07
         ar1    intercept        OTU_5        OTU_3        OTU_6        OTU_2
0.0060247121 0.0003436125 0.0051234368 0.0611676252 0.3401509376 0.6691459691
       OTU_1        OTU_4       OTU_25
0.0035833187 0.3001925971 0.0113659976

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept     OTU_5   OTU_3   OTU_6   OTU_2    OTU_1    OTU_4   OTU_25
      0.6312   365.9873  -1.2659  3.0730  1.6681  0.2000  -0.4769  -1.1963  -4.6991
s.e.  0.2298   102.2315   0.4523  1.6413  1.7487  0.4681   0.1637   1.1547   1.8564

sigma^2 estimated as 147.7:  log likelihood = -102.09,  aic = 224.17
         ar1    intercept        OTU_5        OTU_3        OTU_6        OTU_1
2.070487e-03 5.156367e-05 5.621236e-03 5.870366e-02 3.525421e-01 1.214862e-03
       OTU_4       OTU_25
3.616915e-01 1.220687e-02

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept     OTU_5   OTU_3   OTU_6    OTU_1    OTU_4   OTU_25
      0.6584   383.0450  -1.2267  3.0773  1.6298  -0.5025  -0.9770  -4.6321
s.e.  0.2138    94.6158   0.4430  1.6278  1.7531   0.1553   1.0711   1.8483

sigma^2 estimated as 148.4:  log likelihood = -102.18,  aic = 222.36
         ar1    intercept        OTU_5        OTU_3        OTU_6        OTU_1
5.721428e-06 1.092331e-04 5.635981e-03 8.270531e-02 3.252044e-01 2.741802e-03
      OTU_25
6.197321e-03

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept     OTU_5   OTU_3   OTU_6    OTU_1   OTU_25
      0.7288   351.0559  -1.2328  2.6921  1.7376  -0.4805  -4.9266
s.e.  0.1606    90.7330   0.4453  1.5514  1.7662   0.1604   1.7999

sigma^2 estimated as 152.1:  log likelihood = -102.59,  aic = 221.17
         ar1    intercept        OTU_5        OTU_3        OTU_1       OTU_25
4.547830e-05 5.416774e-05 7.014539e-03 1.580081e-01 2.485879e-03 1.959789e-03

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept     OTU_5   OTU_3    OTU_1   OTU_25
      0.7037   367.4677  -1.2278  1.9487  -0.4862  -3.5105
s.e.  0.1726    91.0278   0.4554  1.3803   0.1607   1.1338

sigma^2 estimated as 158.1:  log likelihood = -103.06,  aic = 220.12
         ar1    intercept        OTU_5        OTU_1       OTU_25
1.155575e-05 2.948192e-06 1.578257e-02 3.115730e-04 4.099725e-03
```

```
Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept     OTU_5     OTU_1    OTU_25
      0.7175   410.7687  -1.1238  -0.5675  -2.3673
s.e.  0.1636    87.8763   0.4656   0.1574   0.8247

sigma^2 estimated as 170.1:  log likelihood = -104.03,  aic = 220.06
```

Abundansen af OTU 24 kan beskrives ved $x_{10t} = 410,77 + 0,72x_{10t-1} - 1,12x_{3t-1} - 0,57x_{7t-1} - 2,37x_{13t-1}$ Af figur 66 ses det, at residualerne tilnærmelsesvis følger en normalfordeling.

```
1 T14_OTU24=modellering(T14_data_matrix,T14_B,10)[length(modellering(T14_data_matrix,T14_B,10))
      -1]
2 T14_OTU24_res=T14_OTU24[[1]]$residuals[-1]
3 par(mfrow=c(1,1))
4 qqnorm(T14_OTU24_res,main=" "); qqline(T14_OTU24_res, main=" ")
```



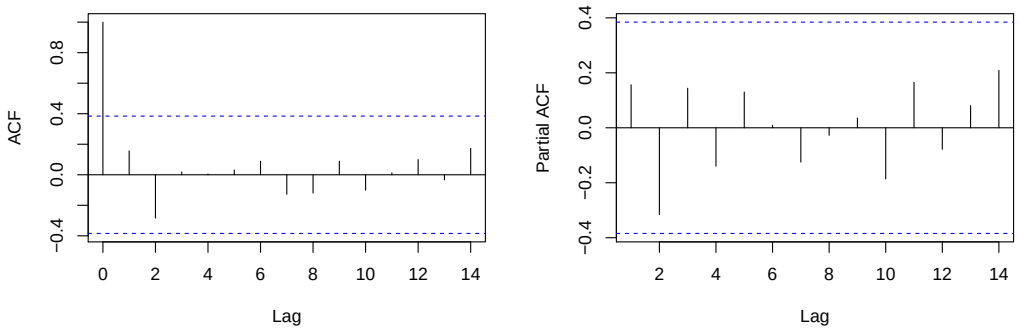
Figur 66: QQ-plot af residualerne for OTU 24.

```
1 Box.test(T14_OTU24_res)
```

```
  Box-Pierce test

data:  T14_OTU24_res
X-squared = 0.63508, df = 1, p-value = 0.4255
```

Af Box-Pierce testen kan man ikke forkaste, at der ingen seriel korrelation er.

```
1 acf(T14_OTU24_res, main=" "); pacf(T14_OTU24_res, main=" ")
```



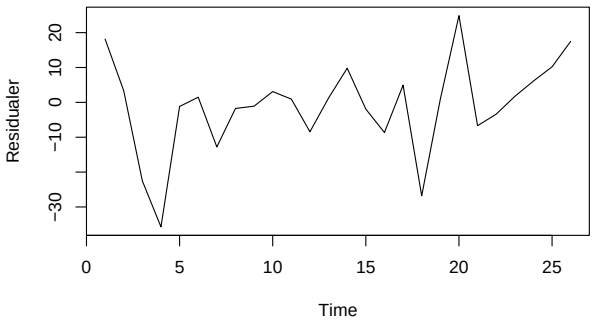
Figur 67: ACF og PACF korrelogrammer for OTU 13.

```
adf.test(T14_OTU24_res,k=0);adf.test(T14_data_matrix[,10],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU24_res
Dickey-Fuller = -4.8206, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(T14_OTU24_res, k = 0) : p-value smaller than printed p-value

  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 10]
Dickey-Fuller = -1.2859, Lag order = 0, p-value = 0.8451
alternative hypothesis: stationary
```

Dickey-Fuller testene viser, at residualerne er stationær, men tidsserien er ikke stationer.

```
plot.ts(T14_OTU24_res,ylab="Residualer")
```



Figur 68: Tidsserie af residualerne for OTU 24.

OTU 7

```
modellering(T14_data_matrix,T14_B,11)
```

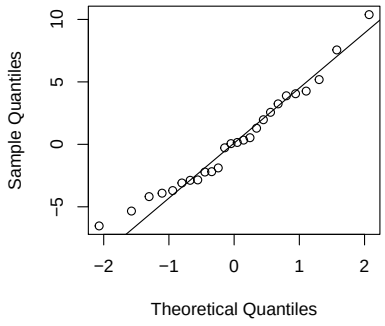
```
        ar1   intercept       OTU_5       OTU_6       OTU_1
0.135944125 0.003334239 0.144590862 0.278221382 0.018551556
```

```
Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_5    OTU_6    OTU_1
      0.3686    51.0865  0.2356  -0.3120  -0.0780
s.e.  0.2472    17.4053  0.1615   0.2877   0.0331

sigma^2 estimated as 14.9:  log likelihood = -72.08,  aic = 156.17
        ar1   intercept       OTU_5       OTU_1
0.028006195 0.005813873 0.218629866 0.036434269

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_5    OTU_1
      0.4905    40.6046  0.1897  -0.0793
s.e.  0.2232    14.7220  0.1542   0.0379

sigma^2 estimated as 15.41:  log likelihood = -72.58,  aic = 155.17
                      ar1                 intercept data[c(NA, 1:26), omgang]
             1.995528e-03              3.627896e-19              1.971629e-02

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  data[c(NA, 1:26), omgang]
      0.5942    58.0817                    -0.0969
s.e.  0.1923     6.4912                     0.0415

sigma^2 estimated as 16.22:  log likelihood = -73.33,  aic = 154.66
```

Abundansen af OTU 7 kan beskrives ved $x_{11t} = 58,08 + 0,59x_{11t-1} - 0,10x_{7t-1} + w_t$

Af figur 69 ses det, at residualerne tilnærmelsesvis følger en normalfordeling.

```
1 T14_OTU7=modellering(T14_data_matrix,T14_B,11)[length(modellering(T14_data_matrix,T14_B,11))-1]
2 T14_OTU7_res=T14_OTU7[[1]]$residuals[-1]
3 qqnorm(T14_OTU7_res,main=" "); qqline(T14_OTU7_res, main=" ")
```



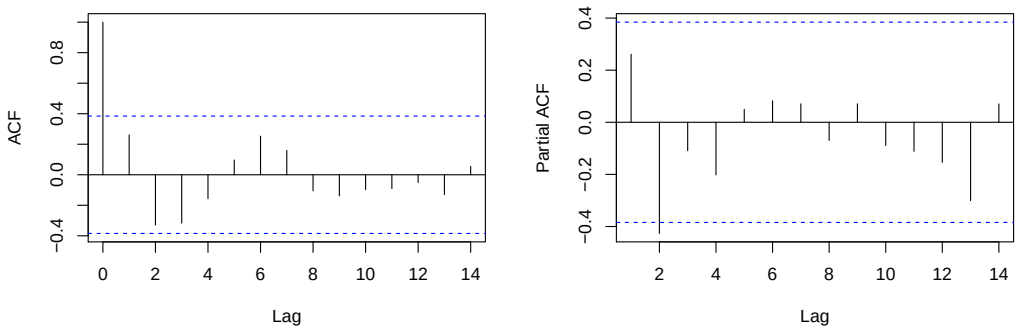
Figur 69: QQ-plot af residualerne for OTU 7.

```
1 Box.test(T14_OTU7_res)
```

```
  Box-Pierce test

data:  T14_OTU7_res
X-squared = 1.7728, df = 1, p-value = 0.183
```

Af Box-Pierce testen kan man ikke forkaste, at der ingen seriel korrelation er.

```
1 acf(T14_OTU7_res, main=" "); pacf(T14_OTU7_res, main=" ")
```



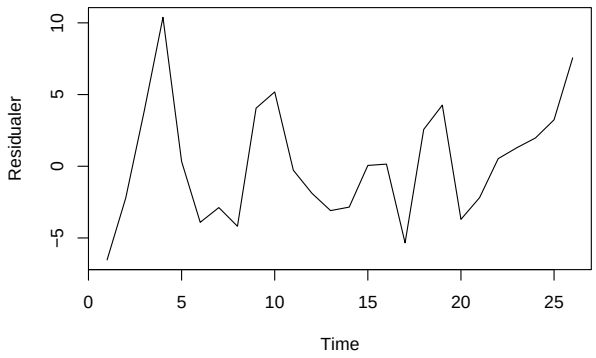
Figur 70: ACF og PACF korrelogrammer for OTU 7.

```
adf.test(T14_OTU7_res,k=0);adf.test(T14_data_matrix[,11],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU7_res
Dickey-Fuller = -3.3995, Lag order = 0, p-value = 0.07785
alternative hypothesis: stationary


  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 11]
Dickey-Fuller = -1.9434, Lag order = 0, p-value = 0.5941
alternative hypothesis: stationary
```

Dickey-Fuller testene viser, at tidsserien ikke er stationær.

```
plot.ts(T14_OTU7_res,ylab="Residualer")
```



Figur 71: Tidsserie af residualerne for OTU 7.

```
adf.test(T14_OTU7[[1]]$residuals[-c(1,27)],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU7[[1]]$residuals[-c(1, 27)]
Dickey-Fuller = -3.746, Lag order = 0, p-value = 0.03957
alternative hypothesis: stationary
```

Ved at fjerne en observation bliver residualprocessen signifikant stationær.

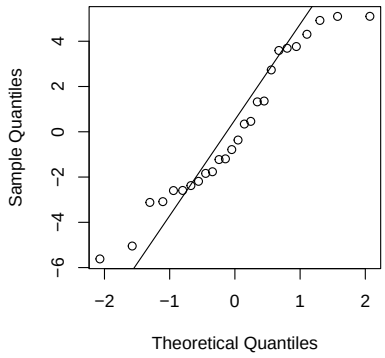
OTU 1406

```
modellering(T14_data_matrix,T14_B,12)
```

```
         ar1    intercept        OTU_1       OTU_13       OTU_15
6.295090e-01 8.448321e-31 4.057380e-01 4.085165e-03 3.033599e-01

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept   OTU_1   OTU_13   OTU_15
      0.0975    61.6601  0.0165  -0.1840  -0.0519
s.e.  0.2020     5.3439  0.0198   0.0641   0.0504

sigma^2 estimated as 6.013:  log likelihood = -60.22,  aic = 132.44
          ar1     intercept        OTU_13        OTU_15
 6.505708e-01 4.092400e-136  1.819621e-03  1.512394e-01

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept   OTU_13   OTU_15
      0.0918    65.5555  -0.1963  -0.0677
s.e.  0.2026     2.6400   0.0629   0.0472

sigma^2 estimated as 6.176:  log likelihood = -60.57,  aic = 131.13
                      ar1                intercept data[c(NA, 1:26), omgang]
             4.654529e-01             1.237374e-140              3.057999e-05

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  data[c(NA, 1:26), omgang]
      0.1485    63.7567                    -0.2433
s.e.  0.2035     2.5254                     0.0584

sigma^2 estimated as 6.643:  log likelihood = -61.52,  aic = 131.04
        ar1   intercept
0.009136142 0.000000000

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept
      0.4389    53.3986
s.e.  0.1684     1.0368

sigma^2 estimated as 9.648:  log likelihood = -69.02,  aic = 144.04
```

Abundansen af OTU 1406 er bestemt til $x_{12t} = 53,40 + 0,44x_{12t-1} + w_t$. Af figur 72 ses det, at residualerne tilnærmelsesvis følger en normalfordeling.

```
T14_OTU1406=modellering(T14_data_matrix,T14_B,12)[length(modellering(T14_data_matrix,T14_B,12))
    -1]
T14_OTU1406_res=T14_OTU1406[[1]]$residuals[-1]
qqnorm(T14_OTU1406_res,main=" "); qqline(T14_OTU1406_res, main=" ")
```



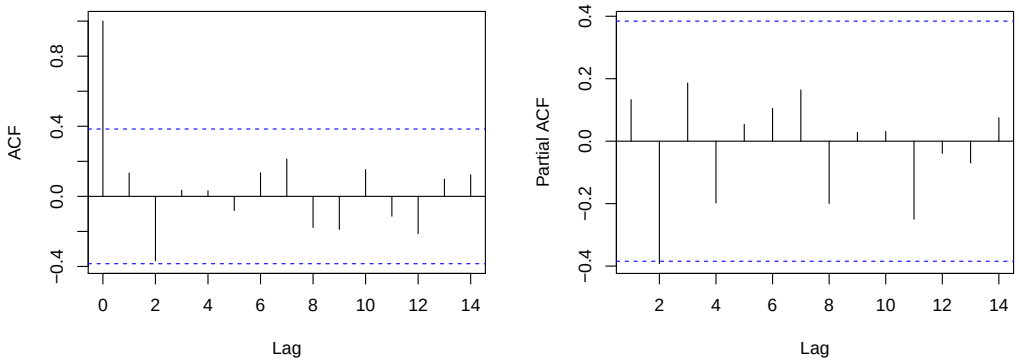
Figur 72: QQ-plot af residualerne for OTU 1406.

```
1 Box.test(T14_OTU1406_res)
```

```
  Box-Pierce test

data:  T14_OTU1406_res
X-squared = 0.45785, df = 1, p-value = 0.4986
```

Af Box-Pierce testen kan man ikke forkaste, at der ingen seriel korrelation er.

```
1 acf(T14_OTU1406_res, main=" "); pacf(T14_OTU1406_res, main=" ")
```



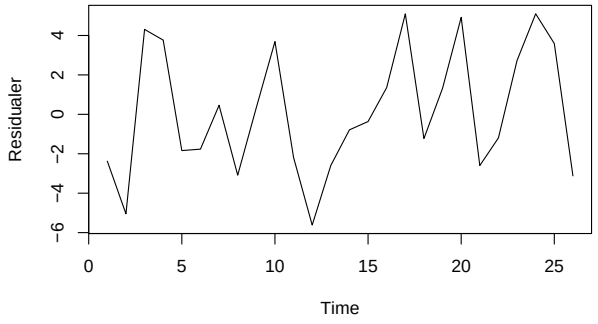
Figur 73: ACF og PACF korrelogrammerne af residualerne for OTU 1406.

```
1 adf.test(T14_OTU1406_res,k=0);adf.test(T14_data_matrix[,12],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU1406_res
Dickey-Fuller = -4.1626, Lag order = 0, p-value = 0.01758
alternative hypothesis: stationary


  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 12]
Dickey-Fuller = -3.4171, Lag order = 0, p-value = 0.07496
alternative hypothesis: stationary
```

Dickey-Fuller testene viser, at residualprocessen er stationær, men da observationerne ikke stationær, er tidsserien ikke er stationær.

```
plot.ts(T14_OTU1406_res,ylab="Residualer")
```



Figur 74: Tidsserie af residualerne for OTU 24.

OTU 25

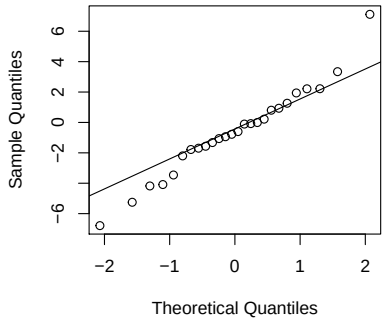
```
modellering(T14_data_matrix,T14_B,13)
```

```
        ar1     intercept         OTU_6        OTU_13        OTU_15
1.768078e-01 3.852167e-16 3.508266e-02 3.275754e-01 2.930553e-01

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_6   OTU_13   OTU_15
      0.2688    68.9659  0.3492  -0.0690  -0.0580
s.e.  0.1990     8.4692  0.1657   0.0704   0.0551

sigma^2 estimated as 5.709:  log likelihood = -59.58,  aic = 131.15
         ar1     intercept         OTU_6        OTU_15
8.066751e-02 3.467921e-14 5.935903e-02 9.198797e-02

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_6   OTU_15
      0.3333    68.3577  0.3340  -0.0868
s.e.  0.1908     9.0187  0.1772   0.0515

sigma^2 estimated as 5.895:  log likelihood = -60.01,  aic = 130.03
                      ar1                  intercept data[c(NA, 1:26), omgang]
             3.519204e-02               2.148216e-12              7.809199e-02

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept  data[c(NA, 1:26), omgang]
      0.3959    63.0495                     0.3445
s.e.  0.1880     8.9756                     0.1956

sigma^2 estimated as 6.51:  log likelihood = -61.33,  aic = 130.66
        ar1   intercept
0.002448919 0.000000000

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept
      0.5701    79.6080
s.e.  0.1882     1.4151
```

```
sigma^2 estimated as 10.57:  log likelihood = -70.34,  aic = 146.68
```

Abundansen af OTU 25 er bestemt til $x_{13t} = 79,61 + 0,57x_{13t-1} + w_t$. Af figur 75 ses det, at residualerne tilnærmelsesvis følger en normalfordeling.

```
1 T14_OTU25=modellering(T14_data_matrix,T14_B,13)[length(modellering(T14_data_matrix,T14_B,13))
      -1]
2 T14_OTU25_res=T14_OTU25[[1]]$residuals[-1]
3 par(mfrow=c(1,1))
4 qqnorm(T14_OTU25_res,main=" "); qqline(T14_OTU25_res, main=" ")
```



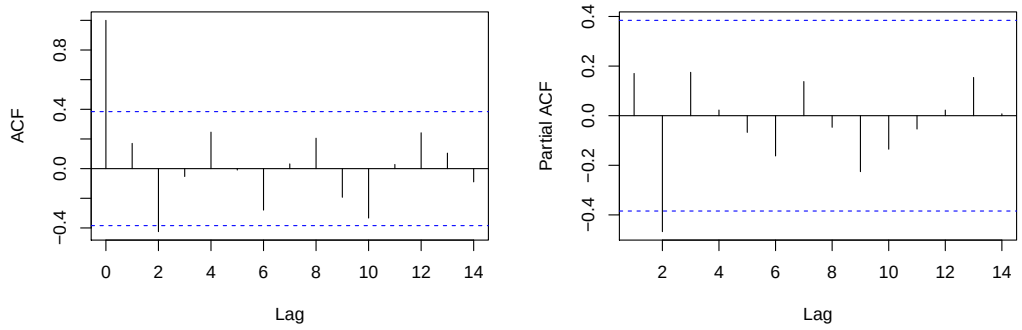
Figur 75: QQ-plot af residualerne for OTU 25.

```
1 Box.test(T14_OTU25_res)
```

```
  Box-Pierce test

data:  T14_OTU25_res
X-squared = 0.75373, df = 1, p-value = 0.3853
```

Af Box-Pierce testen kan man ikke forkaste, at der ingen seriel korrelation er.

```
1 acf(T14_OTU7_res, main=" "); pacf(T14_OTU7_res, main=" ")
```



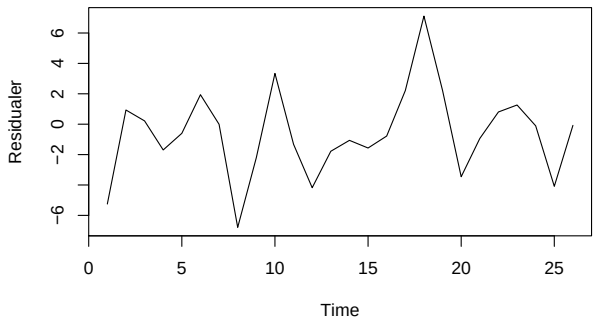
Figur 76: ACF- og PACF korrelogram af residualerne for OTU 25.

```
1 adf.test(T14_OTU25_res,k=0);adf.test(T14_data_matrix[,13],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU25_res
Dickey-Fuller = -4.1582, Lag order = 0, p-value = 0.01774
alternative hypothesis: stationary


  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 13]
Dickey-Fuller = -3.773, Lag order = 0, p-value = 0.03729
alternative hypothesis: stationary
```

Dickey-Fuller testene viser, at tidsserien er stationær.

```
1 plot.ts(T14_OTU25_res,ylab="Residualer")
```



Figur 77: Tidsserie af residualerne for OTU 25.

OTU 15

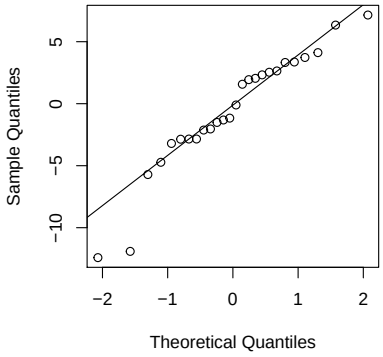
```
1 modellering(T14_data_matrix,T14_B,14)
```

```
         ar1    intercept        OTU_5        OTU_3        OTU_2        OTU_4
7.076840e-38 2.019608e-01 2.040518e-01 4.038699e-02 8.299120e-03 3.932001e-01

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_5   OTU_3    OTU_2   OTU_4
      0.9055    33.7822  0.1920  0.7562  -0.3888  0.3625
s.e.  0.0704    26.4754  0.1512  0.3689   0.1473  0.4246

sigma^2 estimated as 20.4:  log likelihood = -76.95,  aic = 167.9
         ar1    intercept        OTU_5        OTU_3        OTU_2
7.293327e-49 5.360446e-02 2.350666e-01 1.104109e-02 1.265464e-02

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_5   OTU_3    OTU_2
      0.9182    45.0940  0.1798  0.8747  -0.3366
s.e.  0.0625    23.3645  0.1514  0.3442   0.1350

sigma^2 estimated as 20.87:  log likelihood = -77.32,  aic = 166.63
         ar1    intercept        OTU_3        OTU_2
3.783803e-45 4.111202e-02 9.953993e-03 2.901957e-02

Call:
arima(x = data[, x], order = c(1, 0, 0), xreg = data[c(NA, 1:26), omgang])

Coefficients:
         ar1  intercept    OTU_3    OTU_2
      0.9120    48.3905  0.9110  -0.2915
```

```
s.e.  0.0647     23.6930  0.3534    0.1335

sigma^2 estimated as 22.05:  log likelihood = -78,  aic = 166
```

Abundansen af OTU 15 kan beskrives som $x_{14t} = 48,39 + 0,91x_{14t-1} + 0,91x_{4t-1} - 0,29x_{6t-1} + w_t$. Af figur 78 ses det, at residualerne tilnærmelsesvis følger en normalfordeling (være sikker).

```
1 T14_OTU15=modellering(T14_data_matrix,T14_B,14)[length(modellering(T14_data_matrix,T14_B,14))
      -1]
2 T14_OTU15_res=T14_OTU15[[1]]$residuals[-1]
3 par(mfrow=c(1,1))
4 qqnorm(T14_OTU15_res,main=" "); qqline(T14_OTU15_res, main=" ")
```

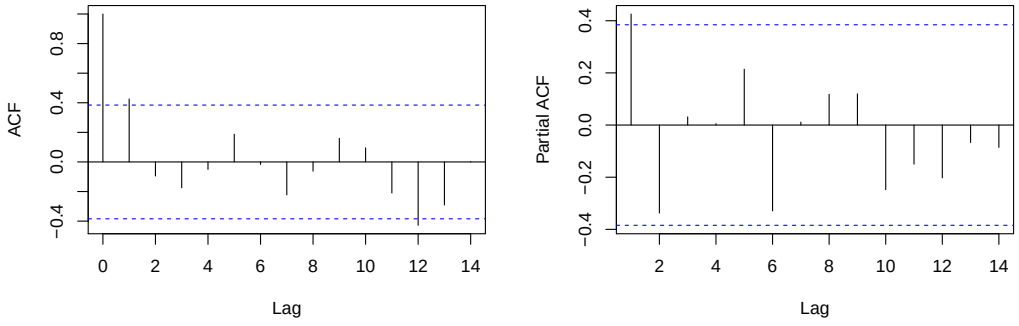


Figur 78: QQ-plot af residualerne for OTU 15.

```
1 Box.test(T14_OTU15_res)
```

```
  Box-Pierce test

data:  T14_OTU15_res
X-squared = 4.7128, df = 1, p-value = 0.02994
```

Af Box-Pierce testen kan man forkaste nuhypotesen om, at der ingen seriel korrelation er. Det vil sige, at der er en signifikant seriel korrelation. Dette understøttes også af ACF og PACF korrelogrammerne på figur 79.



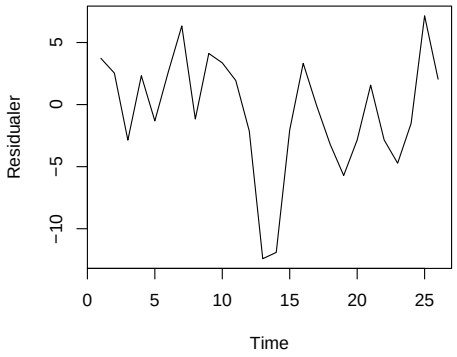
Figur 79: ACF- og PACF korrelogram af residualerne for OTU 15.

```
adf.test(T14_OTU15_res,k=0);adf.test(T14_data_matrix[,14],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU15_res
Dickey-Fuller = -2.956, Lag order = 0, p-value = 0.2082
alternative hypothesis: stationary


  Augmented Dickey-Fuller Test

data:  T14_data_matrix[, 14]
Dickey-Fuller = -1.3807, Lag order = 0, p-value = 0.8089
alternative hypothesis: stationary
```

Dickey-Fuller testene viser, at tidsserien ikke er stationær.

```
plot.ts(T14_OTU15_res,ylab="Residualer")
```



Figur 80: Tidsserie af residualerne for OTU 15.

```
adf.test(T14_OTU15[[1]]$residuals[-c(1,15)],k=0)
```

```
  Augmented Dickey-Fuller Test

data:  T14_OTU15[[1]]$residuals[-c(1, 15)]
Dickey-Fuller = -3.63, Lag order = 0, p-value = 0.04786
alternative hypothesis: stationary
```

Ved at fjerne en af residualerne viser Dickey-Fuller testen at residualprocessen er signifikant stationær. Der konstrueres en overgangsmatrice med tilhørende intercept for VAR(1) baseret på de estimerede univariate tidsserier.

```
Intercept_A_ny=c(rep(0,14))
tmp=c(T14_OTU26,T14_OTU12,T14_OTU3,T14_OTU6,T14_OTU2,T14_OTU1,T14_OTU13,T14_OTU4,T14_OTU24,T14_
    OTU7,T14_OTU1406,T14_OTU25,T14_OTU15)
for(x in c(1,2)){
  Intercept_A_ny[x]=tmp[x][[1]]$coef[2]
}
for(x in c(3:13)){
  Intercept_A_ny[x+1]=tmp[x][[1]]$coef[2]
}
A_ny=matrix( rep( 0, len=14), nrow = 14, ncol=14)
A_ny[1,1]=T14_OTU26[[1]]$coef[1] #OTU26
A_ny[2,2]=T14_OTU12[[1]]$coef[1]; A_ny[2,7]=T14_OTU12[[1]]$coef[3]; A_ny[2,7]=T14_OTU12[[1]]$
    coef[4]#OTU12
#otu5 kan ikke beskrives som en AR(1)
A_ny[4,4]=T14_OTU3[[1]]$coef[1]
A_ny[5,5]=T14_OTU6[[1]]$coef[1]
A_ny[6,6]=T14_OTU2[[1]]$coef[1]; A_ny[6,10]=T14_OTU2[[1]]$coef[3]
A_ny[7,7]=T14_OTU1[[1]]$coef[1]; A_ny[7,10]=T14_OTU1[[1]]$coef[3]; A_ny[7,11]=T14_OTU1[[1]]$
    coef[4]
```

```
A_ny[8,8]=T14_OTU13[[1]]$coef[1];A_ny[8,11]=T14_OTU13[[1]]$coef[3]
A_ny[9,9]=T14_OTU4[[1]]$coef[1];A_ny[9,10]=T14_OTU4[[1]]$coef[3]
A_ny[10,10]=T14_OTU24[[1]]$coef[1];A_ny[10,3]=T14_OTU24[[1]]$coef[3]; A_ny[10,7]=T14_OTU24[[1]]
    $coef[4];A_ny[10,13]=T14_OTU24[[1]]$coef[5]
A_ny[11,11]=T14_OTU7[[1]]$coef[1];A_ny[11,7]=T14_OTU7[[1]]$coef[3];
A_ny[12,12]=T14_OTU1406[[1]]$coef[1]
A_ny[13,13]=T14_OTU25[[1]]$coef[1]
A_ny[14,14]=T14_OTU15[[1]]$coef[1]
```

Residualerne for VAR(1) gemmes i en dataframe.

```
save(Intercept_A_ny, file = "Intercept_A_ny.Rda")
save(A_ny, file = "A_ny.Rda")
```

G Hjælpesætninger

G.1 Definition: [25] Subgradient af en konveks funktion $f : I \to \mathbb{R}$ i et punkt $\theta_0 \in I$ er en skalar k der opfylder

$$f(\theta) - f(\theta_0) \geq k(\theta - \theta_0), \quad \forall \theta \in I.$$

Mængden af subgradienter i et punkt θ_0 kaldes subdifferentialet og defineres som

$$\partial f(\theta)|_{\theta_0} = \Big[\lim_{\theta \to \theta_0^-} \frac{f(\theta) - f(\theta_0)}{\theta - \theta_0}; \lim_{\theta \to \theta_0^+} \frac{f(\theta) - f(\theta_0)}{\theta - \theta_0} \Big] \qquad \blacklozenge$$